Molecular Evolution of Male Reproductive Genetics

Rettie, Elaine

Award date:
2013

Awarding institution:
University of Bath

Molecular Evolution of Male Reproductive Genetics

Elaine Claire Rettie

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

January 2013

"I may not have gone where I intended to go, but I think I have ended up where I intended to be."

- Douglas Adams -

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my supervisor Dr. Steve Dorus, without his patience, advice and occasional pep talks this project would not have been completed, and without his guidance in steering me away from tangents this thesis would be much longer.

To my husband, Jon Rettie, who has been with me the entire journey, providing hugs, love, IT support, general handholding and lots of cups of tea.

To Dorothy and Thomas Mayne without whose love and support, and their immeasurable kindness in taking me and my sister into their home and into their family, I would not have made it to university.

To my sister, Tina Wilkin, for always being on the end of a phone with a story to distract me from thesis-related stress.

To Marios Richards and Dr. Kirril Borziak for sharing their computer skills.

To Dr. Rosie Wasbrough, Dr. Catherine Pink, Alex Ball and Weihoa Zhong for all of their help, support and motivation.

And finally to my RPG group, who by allowing me to pretend to be someone else allowed me to keep my sanity.

Abstract

Spermatozoa in different species are morphologically and physiologically distinct, as well as under different post-copulatory selection regimes. Here we used published data on sperm cell proteomics and microarray expression profiles of the testis and sequential spermatogenic stages to elucidate trends in the evolution of male-related genes. A comparative proteomic analysis of the *Mus musculus* and *Drosophila melanogaster* sperm proteomes demonstrated the conservation of the functional composition of sperm proteomes. Despite this similarity, spermatozoa are known to be a rapidly evolving, highly species specific, cell type. One possible hypothesis that may explain the dichotomy of rapid evolution and conservative constraint is gene duplication. Consistently, a survey of retrotransposition, a RNA-based form of gene duplication, in mammals and *D. melanogaster* revealed that ~20% of known retrogenes encode novel sperm proteins. Further analysis demonstrated that retrotransposition has an important role in the generation of novel sperm genes that function in metabolism. Of particular interest was the observation that sperm retrogenes in mammals and Drosophila were enriched with functions in disparate metabolic pathways, which mirrored the different pathways underlying processes subject to post-copulatory sexual selection due to their role in sperm competition. In addition to this important role in the evolution of sperm there is a general, documented, trend that retrogene are expressed in the testis. However, little is known about the selection on retrogenes as they initially acquire their ability to be expressed. One hypothesis is that the region into which a retrogene is inserted has an especial influence on the evolution of its expression. To determine whether there is an observable effect of retrogene location on retrogene expression a new model for the identification of genomic regions enriched for genes expressed in the testis was developed. Utilising this method significant co-localisation of testis-expressed genes in the *D. melanogaster* genome was observed, together with a significant association between retrogene residence in these genomic regions and the acquisition of retrogene testis expression.

List of Abbreviations

General Abbreviations

ACP	Accessory Gland Protein
BP	Biological Processes
ChIP	Chromatin Immunoprecipitation
GO	Gene Ontology
MS	Mass Spectrometry
MSCI	Meiotic Sex Chromosome Inactivation
Mya	Millions of Years Ago
TE	Transposable Element

Sperm Proteome Abbreviations

DmSP	*Drosophila melanogaster* Sperm Proteome
MmSP	*Mus musculus* Sperm Proteome

Model Abbreviations

C	Characteristic of interest
D	Minimum density of genes with *C* within a neighbourhood
N	Minimum number of genes with *C* within a neighbourhood
NIM	Neighbourhood Identification Model
T2	2-fold higher expression in the testis than the remainder of the organism
T5	5-fold higher expression in the testis than the remainder of the organism
T10	10-fold higher expression in the testis than the remainder of the organism
TE	Expression enriched in the testis
TES	Expression only enriched within the testis
TP	Expression present within the testis
TPS	Expression only present within the testis

Chapter 1

Introduction

The evolution of male reproductive genes is of interest due to the essential role of spermatozoa in fertilisation in sexually reproducing organisms, the extensive morphological and physiological diversity of spermatozoa (Pitnick, Hosken, and Birkhead, 2009), as well as observations that reproductive proteins evolve rapidly (reviewed in Swanson and Vacquier 2002) and that many newly arisen genes are male-biased in their expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004; Marques et al. 2005; Kaessmann 2010). Understanding the evolution of sperm and male genes will provide information into the proteins involved in fertilisation and insights into infertility. Recent advances in sperm cell proteomics have for the first time provide catalogues of the majority of genes that encode protein components of sperm, allowing us to begin to address central questions about their evolution (Karr 2007; Oliva, de Mateo, and Estanyol 2009; Findlay and Swanson 2010). Meanwhile techniques in studying gene expression are providing information into the genes involved in spermatogenesis (Guo et al. 2004; Shima et al. 2004; Chalmel et al. 2007) and genes that change expression in the female reproductive tract in response to sperm (Fazeli et al. 2004; Mcgraw et al. 2004; Prokupek et al. 2009; Bono et al. 2011), providing insights into the obstacles that sperm need to overcome in order for fertilisation to occur. This thesis sets out to identify general trends in the evolution of male-related genes. We demonstrate how the broad conservation of sperm protein composition between species may be offset by the important contribution of retrotransposition to the evolution of sperm novelty in processes under sexual selection, and how retrogenes expression is itself under selection from its genomic location.

1.1 Spermatozoa

Once thought to act solely as delivery vehicles for paternal DNA it is now understood that sperm RNA and proteins are also contributed to the oocyte, and that the epigenome of the sperm may have important, inheritable, implications for the fitness of the embryo (Daxinger and Whitelaw, 2012). Mutations in sperm proteins can lead to defects in the developing offspring and male infertility, either through inability to traverse the female reproductive tract or inability to recognise and fuse with the oocyte. For instance in humans (*Homo sapiens*), mutations in the gene *DEFB126* have been linked with sub-normal fertility (Tollner et al. 2011). *DEFB126* coats spermatozoa in the epididymis and has been demonstrated to inhibit female immune

recognition, effectively cloaking the sperm (Yudin et al. 2005), and to modulate sperm penetration of the cervical mucus and entry to the oviduct (Tollner et al. 2008). While consequences for offspring due to mutations in sperm proteins can be observed in *Drosophila melanogaster* males with mutations in *k81*. These males produce sperm that are able to fertilise eggs, but the embryos fail to develop due to incomplete segregation of the chromatids during the first cellular division (Loppin et al. 2005; Dubruille et al. 2010). In order to understand the molecular causes of male infertility it is important to understand sperm protein content, the process by which spermatozoa are formed and how these may be disrupted.

Despite sharing this essential function in fertilisation spermatozoa are an incredibly diverse cell type. This diversity is believed to be the result of sexual selection mediated by sperm competition (where sperm from two or more different males compete to fertilise a single oocyte) and, in internally fertilising species, co-evolution with the female reproductive tract (Pitnick, Hosken, and Birkhead 2009; Gage 2012). Males from different species have responded to post-copulatory selection in different ways to enhance their reproductive fitness. In the genus Drosophila, sperm gigantism is common (Karr and Pitnick 1996) and sperm length has been correlated with female reproductive tract architecture (Miller and Pitnick 2002). In *D. melanogaster*, longer sperm have greater fertilisation success compared to shorter sperm (Miller and Pitnick 2002). Similarly in the hermaphroditic flatworm genus Macrostomum, spermatozoa from some species have developed stiff bristles that may act as a defensive mechanism to resist their removal by their mating partner after copulation (Schärer et al. 2004 in Higginson et al. 2012). Meanwhile in the frog *Crinia georgiana* (Dziminski et al. 2009), Atlantic salmon (*Salmo salar*) (Gage et al. 2004) and Iberian red deer (*Cervus elaphus hispanicus*) (Malo et al. 2005) fertilisation success is, at least partly, determined by sperm velocity. In the wood mouse (*Apodemus sylvaticus*) spermatozoa have evolved an apical hook, to allow them to join into long trains that improve sperm swimming speed and hence fertilisation success (Moore et al. 2002). These differences may be reflected by disparities in the protein composition of sperm from different species. Therefore understanding of these differences requires in-depth knowledge about the genes involved in spermatogenesis and in providing the proteins to mature sperm.

1.1.1 Transcriptomics and proteomics in the study of sperm

To understand sperm the process by which sperm are formed must also be understood. Spermatogenesis is a highly regulated process in which sperm undergo extensive changes that drastically differentiate them from somatic cells. The process of meiosis (McKee, Yan, and Tsai 2012), the replacement of histones with protamines and further genome compaction (Rathke et al. 2007) and extensive morphological changes such as sperm individualisation and flagellum development (Cheng and Mruk 2010; White-Cooper and Bausek 2010; Fabian and

Brill 2012) mark the most striking differences. The use of transcriptomics is a key component in the study of spermatogenesis and understanding the complexity of the sperm cell. To date, studies of gene expression (transcriptomics) in the testis, and from samples enriched with cells at specific spermatogenic stages, have provided information on the genes contributing to, and regulating, spermatogenesis (Shima et al. 2004; Lim, Tarayrah, and Chen 2012). In addition, comparative microarray studies of human and *Rattus norvegicus* (rat) identified a core set of conserved genes in mammalian spermatogenesis (Chalmel et al. 2007). While transcription studies in *Mus musculus* (mouse) demonstrate that rates of sequence evolution differ depending on timing of expression in spermatogenesis, with those expressed later in spermatogenesis having higher rates of protein evolution (Good and Nachman 2005). However, expression in the testes or during spermatogenesis does not guarantee that a transcribed protein has been incorporated into the mature spermatozoa. For instance, only 23 out of 858 sperm proteins identified in mouse sperm show testis specific expression (Baker et al. 2008). Furthermore while 2,079 genes are predominantly expressed in *D. melanogaster* testes (Chintapalli, Wang, and Dow 2007), only 1,108 genes are identified as empirically encoding components of *D. melanogaster* mature sperm (Dorus et al. 2006; Wasbrough et al. 2010). The use of transcriptomics while providing, and continuing to provide, valuable information on spermatogenesis is unable to reliably elucidate the protein composition of sperm, as transcription within sperm is progressively silenced as spermatogenesis progresses (Hecht 1998). As such the development of other techniques including mass spectroscopy (MS) revolutionised the study of sperm protein content (Karr 2007; Oliva, de Mateo, and Estanyol 2009).

MS is able to empirically identify proteins by their mass-to-charge ratio. Compared to other protein identification techniques, such as immunoprecipitation, MS has led to the identification of large numbers of sperm proteins. For instance, prior to MS analyses < 20 genes were known to effect sperm function or morphology and only 5 empirically demonstrated to encode proteins from mature *D. melanogaster* sperm, currently the *D. melanogaster* sperm proteome (DmSP) comprises >1000 proteins (Dorus et al. 2006; Wasbrough et al. 2010). Furthermore, whole cells can be directly subjected to MS, although better results are obtained if proteins are first removed from the cell and then separated by electrophoresis (Wasbrough et al. 2010). Spermatozoa are particularly suited to whole cell MS as they are relatively easily purified, relatively biochemically simple, and contain much fewer proteins than other eukaryotic cells (Karr 2007; Oliva, de Mateo, and Estanyol 2009). MS has been performed on sperm from a variety of different species, including human (Baker et al. 2007), mouse (Cao, Gerton, and Moss 2006; Stein et al. 2006; Baker et al. 2008; Dorus et al. 2010; Asano et al. 2010) and *Caenorhabditis elegans* (Chu et al. 2006). By revealing the protein content of spermatozoa it has been possible

to perform evolutionary, and comparative proteomic analyses. Analyses in mouse suggest that generally, sperm proteins evolve relatively slowly with the exception of proteins encoding surface membrane proteins (Dorus et al. 2010, see Appendix I). In *D. melanogaster* sperm proteins also are evolving much more conservatively compared with other reproductive proteins, such as accessory gland proteins (Dorus et al. 2006). These observations are distinctly at odds with the observation that spermatozoa are the most rapidly evolving diverse cell types in sexually reproducing species (Pitnick, Hosken, and Birkhead 2009; Gage 2012). To gain a better insight into this paradox, comparative studies between sperm proteomes from different species must be done. To address this in Chapter 2 we present the first detailed cross-species comparison of a sperm proteome (Rettie and Dorus 2012; see Appendix II).

1.1.2 Gene duplication and the evolution of the sperm proteome

Gene duplication is a mechanism for the generation of biological novelty (Lynch and Conery 2003). By duplicating existing genes newly created genes can, potentially, be released from constraints and/or produce novel functions (Lynch and Conery 2000; Lynch 2002; Jun et al. 2009; Kaessmann 2010). Newly created genes can theoretically have several evolutionary fates: redundancy, subfunctionalization (where the ancestral function is partitioned between the two copies) and neofunctionalization (where one or both copies acquire a new function). The mechanism of gene duplication can have an effect on the evolutionary fate of a newly duplicated gene (Kaessmann, Vinckenbosch, and Long 2009). While DNA-based duplicates can either reside as neighbours of their progenitor, and therefore remain under the same chromatin-based regulation, or can disperse throughout the genome, RNA-based duplicates tend to disperse widely from their parental gene. Further, while DNA-based duplicates have the same intron/exon structure as their parental genes and can be copied along with the original genes regulatory regions, RNA-based duplicates are intronless and are rarely duplicated with regulatory sequences (Kaessmann, Vinckenbosch, and Long 2009). Despite these differences many newly created DNA- and RNA-based gene duplicates have testis expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004; Marques et al. 2005; Kaessmann 2010).

In addition to an observed trend for duplicated genes to acquire male-biased expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004; Marques et al. 2005; Kaessmann 2010), there are documented examples where gene duplication has been responsible for the creation of new genes that have functions in male fitness, spermatogenesis and spermatozoa. This includes the expansion of the sperm leucyl aminopeptidase (S-LAP) gene family. During the evolution of this gene family two genes have been created by tandem gene duplication, while another is the result of retrotransposition (Dorus, Wilkin, and Karr 2011: see Appendix III). The S-LAP genes together constitute the most abundant proteins by mass in *D. melanogaster* sperm (Dorus et al.

2006). Also in *D. melanogaster*, the X-linked gene clusters (tektin and Sdic), created by recurrent tandem duplication events, are known to affect sperm fitness. Sperm from mutant strains of *D. melanogaster* that have fewer copies of *Sdic* do not perform as well as wild type sperm in competition assays (Yeh et al. 2012). While mutations within the tektin genes have been documented to influence the outcomes in sperm competition assays (Greenspan and Clark 2011). Similarly in mammals, the retrotransposed genes *Utp14b* and *Pgk2* have important functions in spermatogenesis (Bradley et al. 2004) and sperm function (Danshina et al. 2010), respectively. Gene duplicates have been frequently observed to resulting from genes in metabolic pathways in both Drosophila (Gallach, Chandrasekaran, and Betrán 2010) and mammals (Vemuganti, De Villena, and O'Brien 2010). Further these gene duplicates are proposed to have functions in sperm. As sperm competitive ability is likely to be under intense sexual selection, we propose that gene duplication may be enriched in those processes that enhance sperm fitness and that this would differ depending on species. The existence of gene duplicates encoding sperm protein components is documented in *D. melanogaster* (Dorus et al. 2008). As such gene duplication may be a solution to the paradox of rapid evolution and concurrent conservation in spermatozoa. In Chapter 3 we present a detailed analysis to determine if retrotransposition is a common mechanism that has contributed proteins, which function in processes related to sperm competitive ability, to both the mouse and Drosophila sperm proteomes.

1.2 Retrotransposition

Retrotransposition creates new genes (retrogenes) by the reverse transcription of a processed mRNA molecule, which is then incorporated into the genome. Due to this process, retrogenes generally have no introns and insert at a distance (often interchromosomally) from their parental gene. Unlike DNA based duplicates, retrogenes are generally not duplicated with the *cis* regulatory sequences of their parental genes. This later characteristic suggests that retrogenes should not be functionally equivalent to their parental copies nor be expressed (Kaessmann, Vinckenbosch, and Long 2009). Many retrogenes however, have expression and function, although it is unclear how they have acquired new regulatory sequences. In particular, retrogenes tend to have testis expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004) and several have documented functions spermatogenesis, such as *Utp14b* in mouse (Bradley et al. 2004) and sperm (Dorus et al. 2008). However, why retrogenes have this tendency towards testis expression and how they become testis expressed are current, unresolved, topics in biology.

1.2.1 Retrogenes: acquiring regulatory sequences

There are several hypotheses of how retrogenes may acquire new *cis* regulation, including *de novo* mutations of nearby genomic sequences to form promoters, co-option of a neighbouring genes sequences, integration into an existing gene or inheritance of the parental regulatory sequences (Kaessmann, Vinckenbosch, and Long 2009). As the mechanism of RNA-duplication requires a processed mRNA molecule, it seems to preclude the inheritance of the source genes promoter unless this region is downstream of the transcriptional start site in which case it may be copied with the remainder of the sequence. However, in a study of human retrogenes, 16 out of 29 retrogenes had promoters that had been copied along with the surrounding coding sequence of the parental gene (Okamura and Nakai 2008). Although a study in *D. melanogaster* demonstrated that retrogenes were generally not duplicated along with regulatory sequences from their parental gene (Bai, Casola, and Betrán, 2008). Alternatively, other retrogenes appear to have become functional by inserting into an existing gene forming chimeric constructs, such as the Drosophila genes *sphinx* (Wang et al. 2002) and *jingwei* (Long and Langley 1993). There is also evidence that retrogenes can co-opt or share the regulatory sequences of neighbouring genes (Vinckenbosch, Dupanloup, and Kaessmann 2006).

1.2.2 Retrogenes and testis neighbourhoods

As retrotransposition must occur within the germline in order to be inherited, and as retrotransposition requires mRNA to occur, parental genes are likely therefore to be expressed in the germline (i.e. the testis) (Kaessmann, Vinckenbosch, and Long 2009). It therefore seems plausible that regions in open and active chromatin formation will be important in retrogene

evolution. To understand how the genomic landscape may influence the expressional evolution of retrogenes, we must first understand genome organisation. Gene order in the eukaryotic genome is not random, and there is increasing evidence that genes are co-localised along chromosomes with respect to their expression (reviewed by Hurst, Pál, and Lercher 2004). In eukaryotes, gene expression is regulated at the individual gene level by *cis*-regulatory sequences; but modifications to the state of chromatin can span many genes (chromatin domains) by controlling access of the transcription proteins to the DNA sequence (Mellor, Dudek, and Clynes 2008; Woodcock and Ghosh 2010; Lelli, Slattery, and Mann 2012). As such, it is likely that there is selection within chromatin domains for genes with similar patterns of expression (Oliver, Parisi, and Clark 2002). There is evidence in mouse and Drosophila that testis expressed genes and genes that encode sperm proteins have co-localised into gene neighbourhoods which are more numerous and larger than expected by random gene order (Boutanaev et al. 2002; Divina et al. 2005; Li, Lee, and Zhang 2005; Dorus et al. 2006; Wasbrough et al. 2010). Furthermore, there is evidence in *D. melanogaster* that these testis gene neighbourhoods are associated with transcriptionally silenced regions of the nucleus in somatic cells but not in testis cells (Shevelyov et al. 2009). As such genomic regions that are enriched for testis-expressed genes likely represent open chromatin regions that may be favourable for retrogene insertion. Furthermore, presence within genome regions enriched with testis-expressed genes may be associated with the testis expression of retrogenes, either due to retention of retrogenes that are testis expressed in order to preserve the integrity of the neighbourhood or due to the evolution of testis expression from co-option and sharing of neighbouring genes regulatory regions. To address these two hypotheses we determined whether there was an over representation of retrogenes in testis gene neighbourhoods and whether there was an association between retrogene residence in testis neighbourhoods and testis expression. However, we first needed a method to assess the extent to which testis genes are spatially co-localised in the genome. Many of the methods developed to identify gene neighbourhoods introduce parameters that may not be biologically realistic, including the setting pre-determined neighbourhood boundaries, and not accounting for asymmetry in gene neighbourhoods. We developed a new flexible algorithm for the identification of gene neighbourhoods (see Chapter 4). This method has been used to assess the co-localisation of genes that encode sperm proteins in *D. melanogaster* (Wasbrough et al. 2010, see Appendix IV), and was successful in identifying genomic regions enriched for testis-expressed genes.

In Drosophila, retrogenes have been observed to reside in regions of the genome that are enriched for genes expressed in the testis, and it is proposed that this may have influenced their acquisition of testis expression (Bai, Casola, and Betrán 2008; Dorus et al. 2008). In *D. melanogaster*, the four flanking genes of each testis retrogenes were examined and no

association between the retrogenes and the neighbours expression in the testis was observed (Bai, Casola, and Betrán 2008). However, studies of gene co-localisation have identified neighbourhoods of testis-expressed genes that span more than four genes (Boutanaev et al. 2002; Divina et al. 2005; Li, Lee, and Zhang 2005; Dorus et al. 2008). It is therefore possible that the region examined by Bai *et al*. (2008) surrounding the retrogene was too small to detect an association between retrogene and neighbours expression. We provide evidence that there is an association between retrogene expression and residence within these larger testis gene neighbourhoods (see Chapter 5).

1.3 Summary

Despite the importance of spermatozoa in fertility, much remains to be studied in terms of their protein content, function and evolution. Advances in MS technology and methodology have provided numerous catalogues of empirically identified reproductive proteins, and have been particularly important in the study of sperm proteomes due to the difficulties in studying this cell type (Oliva, De Mateo, and Estanyol 2009). This dissertation sets out to address the functional similarities of identified proteins in mature spermatozoa, and the role of retrotransposition, in two taxa with spermatozoa under different post-copulatory selection. Our comparative study of the *D. melanogaster* and mouse sperm proteomes demonstrate that despite differences in sperm morphology, physiology and post-copulatory selection, there are considerable parallels in their protein content, in terms of functions. Furthermore, we observe that retrotransposition has contributed to sperm protein evolution in both mammals and Drosophila. Despite the well-studied movement and expression of retrogenes, the numerous case studies in which they have demonstrated importance to sperm function or male-fitness, there remains uncertainty surrounding the genome level processes that effect the initial acquisition of retrogene expression. In addition this dissertation addressed previous problems in the methods used to identify neighbourhoods of similarly expressed genes, via the creation of a novel model, which was subsequently used to investigate the role of retrogene location on retrogene expression. We observed that there was a highly significant association between retrogene expression in the testis and presence within a genomic region enriched for testis genes. Indeed all retrogenes within these testis neighbourhoods have acquired testis expression. We therefore propose that while genome level selection (e.g. chromatin domains) may have significant effects on the initial acquisition of expression in retrogenes, phenotypic level selection (e.g. sperm fitness traits) may further develop retrogene function.

1.4 References

Asano, Atsushi, Jacquelyn L Nelson, Sheng Zhang, and Alexander J Travis. 2010. "Characterization of the Proteomes Associating with Three Distinct Membrane Raft Subtypes in Murine Sperm." Proteomics 10 (19): 3494–3505.

Bai, Yongsheng, Claudio Casola, and Esther Betrán. 2008. "Evolutionary Origin of Regulatory Regions of Retrogenes in Drosophila." BMC Genomics 9: 241.

Baker, Mark A, Louise Hetherington, Gabi M Reeves, and R John Aitken. 2008. "The Mouse Sperm Proteome Characterized via IPG Strip Prefractionation and LC-MS/MS Identification." Proteomics 8 (8): 1720–1730.

Baker, Mark A, Gabi Reeves, Louise Hetherington, Jörg Müller, Inke Baur, and R John Aitken. 2007. "Identification of Gene Products Present in Triton X-100 Soluble and Insoluble Fractions of Human Spermatozoa Lysates Using LC-MS/MS Analysis." *Proteomics. Clinical Applications* 1 (5): 524–532.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. "Retroposed New Genes Out of the X in Drosophila." Genome Research 12 (12): 1854–1859.

Bono, Jeremy M, Luciano M Matzkin, Erin S Kelleher, and Therese A Markow. 2011. "Postmating Transcriptional Changes in Reproductive Tracts of Con- and Heterospecifically Mated Drosophila Mojavensis Females." Proceedings of the National Academy of Sciences of the United States of America 108 (19): 7878–7883.

Boutanaev, Alexander M, Aila I Kalmykova, Yuri Y Shevelyov, and Nurminsky Dmitry I. 2002. "Large Clusters of Co-expressed Genes in the Drosophila Genome." Nature 420 (6916): 666–669.

Bradley, Julie, Andrew Baltus, Helen Skaletsky, Morgan Royce-Tolland, Ken Dewar, and David C Page. 2004. "An X-to-autosome Retrogene Is Required for Spermatogenesis in Mice." Nature Genetics 36 (8): 872–976.

Cao, Wenlei, George L Gerton, and Stuart B Moss. 2006. "Proteomic Profiling of Accessory Structures from the Mouse Sperm Flagellum." Molecular & Cellular Proteomics 5 (5): 801–810.

Chalmel, Frédéric, Antoine D Rolland, Christa Niederhauser-Wiederkehr, Sanny S W Chung, Philippe Demougin, Alexandre Gattiker, James Moore, et al. 2007. "The Conserved Transcriptome in Human and Rodent Male Gametogenesis." Proceedings of the National Academy of Sciences of the United States of America 104 (20): 8346–8351.

Cheng, C Yan, and Dolores D Mruk. 2010. "The Biology of Spermatogenesis: The Past, Present and Future." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1546): 1459–1463.

Chintapalli, Venkateswara R, Jing Wang, and Julian A T Dow. 2007. "Using FlyAtlas to Identify Better Drosophila Melanogaster Models of Human Disease." Nature Genetics 39 (6): 715–720.

Chu, Diana S, Hongbin Liu, Paola Nix, Tammy F Wu, Edward J Ralston, John R Yates, and Barbara J Meyer. 2006. "Sperm Chromatin Proteomics Identifies Evolutionarily Conserved Fertility Factors." Nature 443 (7107): 101–105.

Danshina, Polina V, Christopher B Geyer, Qunsheng Dai, Eugenia H Goulding, William D Willis, G Barrie Kitto, John R McCarrey, E M Eddy, and Deborah A O'Brien. 2010. "Phosphoglycerate Kinase 2 (PGK2) Is Essential for Sperm Function and Male Fertility in Mice." Biology of Reproduction 82 (1): 136–145.

Daxinger, Lucia, and Emma Whitelaw. 2012. "Understanding Transgenerational Epigenetic Inheritance via the Gametes in Mammals." Nature Reviews, Genetics 13 (3): 153–162.

Divina, Petr, Cestmír Vlcek, Petr Strnad, Václav Paces, and Jirí Forejt. 2005. "Global Transcriptome Analysis of the C57BL/6J Mouse Testis by SAGE: Evidence for Nonrandom Gene Order." BMC Genomics 6: 29.

Dorus, Steve, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L Karr. 2006. "Genomic and Functional Evolution of the Drosophila Melanogaster Sperm Proteome." Nature Genetics 38 (12): 1440–1445.

Dorus, Steve, Zoë N Freeman, Elizabeth R Parker, Benjamin D Heath, and Timothy L Karr. 2008. "Recent Origins of Sperm Genes in Drosophila." Molecular Biology and Evolution 25 (10): 2157–2166.

Dorus, Steve, Elizabeth R Wasbrough, Jennifer Busby, Elaine C Wilkin, and Timothy L Karr. 2010. "Sperm Proteomics Reveals Intensified Selection on Mouse Sperm Membrane and Acrosome Genes." Molecular Biology and Evolution 27 (6): 1235–1246.

Dorus, Steve, Elaine C Wilkin, and Timothy L Karr. 2011. "Expansion and Functional Diversification of a Leucyl Aminopeptidase Family That Encodes the Major Protein Constituents of Drosophila Sperm." BMC Genomics 12: 177.

Dubruille, Raphaëlle, Guillermo A Orsi, Lætitia Delabaere, Elisabeth Cortier, Pierre Couble, Gabriel A B Marais, and Benjamin Loppin. 2010. "Specialization of a Drosophila Capping Protein Essential for the Protection of Sperm Telomeres." *Current Biology* 20 (23): 2090–2099.

Dziminski, Martin A, J Dale Roberts, Maxine Beveridge, and Leigh W Simmons. 2009. “Sperm Competitiveness in Frogs: Slow and Steady Wins the Race.” Proceedings of the Royal Society B - Biological Sciences 276 (1675): 3955–3961.

Emerson, J J, Henrik Kaessmann, Esther Betrán, and Manyuan Long. 2004. “Extensive Gene Traffic on the Mammalian X Chromosome.” Science 303 (5657): 537–540.

Fabian, Lacramioara, and Julie A Brill. 2012. “Drosophila Spermiogenesis: Big Things Come from Little Packages.” *Spermatogenesis* 2 (3): 197–212.

Fazeli, Alireza, Nabeel A Affara, Michael Hubank, and William V Holt. 2004. “Sperm-induced Modification of the Oviductal Gene Expression Profile After Natural Insemination in Mice.” Biology of Reproduction 71 (1): 60–65.

Findlay, Geoffrey D, and Willie J Swanson. 2010. “Proteomics Enhances Evolutionary and Functional Analysis of Reproductive Proteins.” BioEssays 32 (1): 26–36.

Gage, Matthew J G. 2012. “Complex Sperm Evolution.” Proceedings of the National Academy of Sciences of the United States of America 109 (12): 4341–4342.

Gage, Matthew J.G., Christopher P. Macfarlane, Sarah Yeates, Richard G. Ward, Jeremy B. Searle, and Geoffrey A. Parker. 2004. “Spermatozoal Traits and Sperm Competition in Atlantic Salmon: Relative Sperm Velocity Is the Primary Determinant of Fertilization Success.” Current Biology 14 (1): 44–47.

Gallach, Miguel, Chitra Chandrasekaran, and Esther Betrán. 2010. “Analyses of Nuclearly Encoded Mitochondrial Genes Suggest Gene Duplication as a Mechanism for Resolving Intralocus Sexually Antagonistic Conflict in Drosophila.” Genome Biology and Evolution 2: 835–850.

Good, Jeffrey M, and Michael W Nachman. 2005. “Rates of Protein Evolution Are Positively Correlated with Developmental Timing of Expression During Mouse Spermatogenesis.” Molecular Biology and Evolution 22 (4): 1044–1052.

Greenspan, Leah, and Andrew G Clark. 2011. “Associations Between Variation in X Chromosome Male Reproductive Genes and Sperm Competitive Ability in Drosophila Melanogaster.” International Journal of Evolutionary Biology 2011: 214280–214289.

Guo, Rui, Zuoren Yu, Jikui Guan, Yehua Ge, Jing Ma, Sai Li, Shali Wang, Shepu Xue, and Daishu Han. 2004. “Stage-specific and Tissue-specific Expression Characteristics of Differentially Expressed Genes During Mouse Spermatogenesis.” Molecular Reproduction and Development 67 (3): 264–272.

Hecht, N B. 1998. “Molecular Mechanisms of Male Germ Cell Differentiation.” BioEssays 20 (7): 555–561.

Higginson, Dawn M, Kelly B Miller, Kari A Segraves, and Scott Pitnick. 2012. “Convergence, Recurrence and Diversification of Complex Sperm Traits in Diving Beetles (Dytiscidae).” Evolution 66 (5): 1650–1661.

Hurst, Laurence D, Csaba Pál, and Martin J Lercher. 2004. “The Evolutionary Dynamics of Eukaryotic Gene Order.” Nature Reviews. Genetics 5 (4): 299–310.

Jun, Jin, Paul Ryvkin, Edward Hemphill, Ion Mandoiu, and Craig Nelson. 2009. “The Birth of New Genes by RNA- and DNA-mediated Duplication During Mammalian Evolution.” Journal of Computational Biology 16 (10): 1429–1444.

Kaessmann, Henrik. 2010. “Origins, Evolution, and Phenotypic Impact of New Genes.” Genome Research 20 (10): 1313–1326.

Kaessmann, Henrik, Nicolas Vinckenbosch, and Manyuan Long. 2009. “RNA-based Gene Duplication: Mechanistic and Evolutionary Insights.” Nature Reviews. Genetics 10 (1): 19–31.

Karr, Timothy L. 2007. “Fruit Flies and the Sperm Proteome.” Human Molecular Genetics 16 (2): 124–133.

Karr, Timothy L, and Scott Pitnick. 1996. “The Ins and Outs of Fertilization.” Nature 379 (6564): 405–406.

Lelli, Katherine M, Matthew Slattery, and Richard S Mann. 2012. “Disentangling the Many Layers of Eukaryotic Transcriptional Regulation.” Annual Review of Genetics 46: 43–68.

Li, Quan, Bernett T K Lee, and Louxin Zhang. 2005. “Genome-scale Analysis of Positional Clustering of Mouse Testis-specific Genes.” BMC Genomics 6: 7.

Lim, Cindy, Lama Tarayrah, and Xin Chen. 2012. “Transcriptional Regulation During Drosophila Spermatogenesis.” Spermatogenesis 2 (3): 158–166.

Long, M, and C H Langley. 1993. “Natural Selection and the Origin of Jingwei, a Chimeric Processed Functional Gene in Drosophila.” Science 260 (5104): 91–95.

Loppin, Benjamin, David Lepetit, Steve Dorus, Pierre Couble, and Timothy L Karr. 2005. “Origin and Neofunctionalization of a Drosophila Paternal Effect Gene Essential for Zygote Viability.” Current Biology 15 (2): 87–93.

Lynch, Michael. 2002. “Gene Duplication and Evolution.” Science 297 (5583): 945–947.

Lynch, Michael, and John S Conery. 2000. “The Evolutionary Fate and Consequences of Duplicate Genes.” Science 290 (5494): 1151–1155.

Lynch, Michael and Conery, John, S. 2003. “The Origins of Genome Complexity.” Science 302 (5649): 1401–1404.

Malo, Aurelio F, J Julián Garde, Ana J Soler, Andrés J García, Montserrat Gomendio, and Eduardo R S Roldan. 2005. “Male Fertility in Natural Populations of Red Deer Is Determined by Sperm Velocity and the Proportion of Normal Spermatozoa.” Biology of Reproduction 72 (4): 822–829.

Marques, Ana Claudia, Isabelle Dupanloup, Nicolas Vinckenbosch, Alexandre Reymond, and Henrik Kaessmann. 2005. “Emergence of Young Human Genes After a Burst of Retroposition in Primates.” PLoS Biology 3 (11): 1970–1979.

Mcgraw, Lisa A, Greg Gibson, Andrew G Clark, and Mariana F Wolfner. 2004. “Genes Regulated by Mating , Sperm , or Seminal Proteins in Mated Female Drosophila Melanogaster.” Current Biology 14 (16): 1509–1514.

McKee, Bruce D, Rihui Yan, and Jui-He Tsai. 2012. “Meiosis in Male Drosophila.” Spermatogenesis 2 (3) : 167–184.

Mellor, Jane, Peter Dudek, and David Clynes. 2008. “A Glimpse into the Epigenetic Landscape of Gene Regulation.” Current Opinion in Genetics & Development 18 (2): 116–122.

Miller, Gary T, and Scott Pitnick. 2002. “Sperm-female Coevolution in Drosophila.” Science 298 (5596): 1230–1233.

Moore, Harry, Katerina Dvorakova, Nicholas Jenkins, and William Breed. 2002. “Exceptional Sperm Cooperation in the Wood Mouse.” Nature 418 (6894): 174–177.

Okamura, Kohji, and Kenta Nakai. 2008. “Retrotransposition as a Source of New Promoters.” Molecular Biology and Evolution 25 (6): 1231–1238.

Oliva, Rafael, Sara de Mateo, and Josep Maria Estanyol. 2009. “Sperm Cell Proteomics.” Proteomics 9 (4): 1004–1017.

Oliver, Brian, Michael Parisi, and David Clark. 2002. “Gene Expression Neighbourhoods.” Journal of Biology 1 (1): 4.

Pitnick, S, DJ Hosken, and TR Birkhead. 2009. “Sperm Morphological Diversity.” In Sperm Biology: An Evolutionary Perspective, ed. Pitnick S Birkhead TR, Hosken DJ, 69–149. USA: Academic Press.

Prokupek, A M, S D Kachman, I Ladunga, and L G Harshman. 2009. “Transcriptional Profiling of the Sperm Storage Organs of Drosophila Melanogaster.” Insect Molecular Biology 18 (4): 465–475.

Rathke, Christina, Willy M Baarends, Sunil Jayaramaiah-Raja, Marek Bartkuhn, Rainer Renkawitz, and Renate Renkawitz-Pohl. 2007. “Transition from a Nucleosome-based to a Protamine-based Chromatin Configuration During Spermiogenesis in Drosophila.” Journal of Cell Science 120 (9): 1689–1700.

Rettie, Elaine C, and Steve Dorus. 2012. “Drosophila Sperm Proteome Evolution - Insights from Comparative Genomic Approaches.” Spermatogenesis 2 (3): 213–223.

Schärer, L., G. Joss, and P. Sandner. 2004. "Mating behaviour of the marine turbellarian *Macrostomum* sp.: these worms suck". Marine Biology 145: 373–380.

Shevelyov, Y Y, S A Lavrov, L M Mikhaylova, I D Nurminsky, R J Kulathinal, K S Egorova, Y M Rozovsky, and D I Nurminsky. 2009. “The B-type Lamin Is Required for Somatic Repression of Testis-specific Gene Clusters.” Proceedings of the National Academy of Sciences of the United States of America 106 (9): 3282–3287.

Shima, James E, Derek J McLean, John R McCarrey, and Michael D Griswold. 2004. “The Murine Testicular Transcriptome: Characterizing Gene Expression in the Testis During the Progression of Spermatogenesis.” Biology of Reproduction 71 (1): 319–330.

Stein, Kathryn K, Jowell C Go, William S Lane, Paul Primakoff, and Diana G Myles. 2006. “Proteomic Analysis of Sperm Regions That Mediate Sperm-egg Interactions.” Proteomics 6 (12): 3533–3543.

Swanson, Willie J, and Victor D Vacquier. 2002. “The Rapid Evolution of Reproductive Proteins." Nature Reviews, Genetics 3 (2): 137–144.

Tollner, Theodore L, Ashley I Yudin, Cathy A Treece, James W Overstreet, and Gary N Cherr. 2008. “Macaque Sperm Coating Protein DEFB126 Facilitates Sperm Penetration of Cervical Mucus.” Human Reproduction 23 (11): 2523–2534.

Tollner, TL, SA Venners, EJ Hollox, AI Yudin, X Liu, and Tang G. 2011. “A Common Mutation in the Defensin DEFB126 Causes Impaired Sperm Function and Subfertility.” Science Translational Medicine 3 (92): 92ra65.

Vemuganti, Soumya A, Fernando Pardo-Manuel de Villena, and Deborah A O’Brien. 2010. “Frequent and Recent Retrotransposition of Orthologous Genes Plays a Role in the Evolution of Sperm Glycolytic Enzymes.” BMC Genomics 11: 285.

Vinckenbosch, Nicolas, Isabelle Dupanloup, and Henrik Kaessmann. 2006. “Evolutionary Fate of Retroposed Gene Copies in the Human Genome.” Proceedings of the National Academy of Sciences of the United States of America 103 (9): 3220–3225.

Wang, Wen, Frédéric G Brunet, Eviatar Nevo, and Manyuan Long. 2002. “Origin of Sphinx, A Young Chimeric RNA Gene in Drosophila Melanogaster.” Proceedings of the National Academy of Sciences of the United States of America 99 (7): 4448–4453.

Wasbrough, Elizabeth R, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L Karr. 2010. “The Drosophila Melanogaster Sperm proteome-II (DmSP-II).” Journal of Proteomics 73 (11): 2171–2185.

White-Cooper, Helen, and Nina Bausek. 2010. “Evolution and Spermatogenesis.” Philosophical Transcations of the Royal Scociety B - Biological Sciences 365 (1546): 1465–1480.

Woodcock, Christopher L, and Rajarshi P Ghosh. 2010. “Chromatin Higher-order Structure and Dynamics.” Cold Spring Harbor Perspectives in Biology 2 (5): a000596.

Yeh, Shu-Dan, Tiffanie Do, Carolus Chan, Adriana Cordova, Francisco Carranza, Eugene A Yamamoto, Mashya Abbassi, et al. 2012. “Functional Evidence That a Recently Evolved Drosophila Sperm-specific Gene Boosts Sperm Competition.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (6): 2043–2048.

Yudin, Ashley I, Suzanne E Generao, Theodore L Tollner, Catherine A Treece, James W Overstreet, and Gary N Cherr. 2005. “Beta-defensin 126 on the Cell Surface Protects Sperm from Immunorecognition and Binding of Anti-sperm Antibodies.” Biology of Reproduction 73 (6): 1243–1252.

Chapter 2

Evolution of the sperm proteome: comparative proteomics analysis of the *Drosophila melanogaster* and *Mus Musculus* sperm proteomes

Primary data adapted from Rettie, E. C. and Dorus, S. (2012). Drosophila sperm proteome evolution: Insights from comparative genomic approaches. *Spermatogenesis* **2**(3): 213-223. Published manuscript provided in Appendix II.

2.1 Introduction

Whilst spermatozoa share the same function, delivering the paternal genetic information to the oocyte, they demonstrate incredible morphological diversity between species (Pitnick, Hosken, and Birkhead 2009). Due to their central role in fertilization, sperm, and the genes underlying sperm form and function, are strong focal points for sexual selection to operate. Selection on sperm is primarily mediated by sperm competition and sperm-female interactions. Sperm competition occurs when sperm from multiple males compete to fertilize a single ovum (Parker 1970). The intensity of sperm competition has been experimentally shown to affect both sperm number and quality (Hosken and Ward, 2001; Gomendio and Roldan, 2008; Kleven et al. 2008; Firman & Simmons, 2011). Morphology can also affect sperm competitive ability; for example, in Drosophila, longer sperm have been experimentally shown to out compete shorter sperm (Miller and Pitnick 2002). Sperm morphology is also closely linked to fertilization success, and infertility can be linked to teratospermia (increase in the number of sperm with morphological defects), as well as asthenozoospermia (sperm with impaired motility) and oligospermia (reduced or no spermatozoa). It is also likely that defects in sperm morphology can be used to assay sperm quality. For instance, human sperm that have "thumbprint-like" vacuoles in the sperm head membrane are associated with a reduction in male fertility (Boitrelle et al. 2011). Further selective pressures on sperm morphology and physiology are imposed through the architecture of, and proteins in, the female reproductive tract (Miller and Pitnick 2002; Holt and Fazeli 2010), as well as sperm-egg interactions (Nixon, Aitken, and McLaughlin 2007; Hasan, Fukami, and Sato 2011). Together with sperm competition, these factors drive the evolution of sperm. Different taxa have different reproductive biologies and different mating systems that modulate the intensity of sperm competition, resulting in different selective drivers and constraints on sperm. The diversity of these selection pressures may explain the observed variation in sperm morphology, while the shared role in fertilization may explain the conservation of essential pathways associated with structure and metabolism.

The advent of mass spectroscopy (MS) has provided a very detailed understanding of which proteins contribute to spermatozoa as well as insights into their evolution (Oliva, De Mateo, and Estanyol 2009). Prior to the development and application of MS, which allows the empirical identification of large numbers of proteins from single samples, the study of protein composition within spermatozoa was difficult. This was primarily due to the progressive transcriptional silencing in sperm as the nucleus is compacted and repackaged (Hecht 1998) rendering it difficult to directly assay gene activity (e.g. microarray analysis or targeted gene insertion). However, spermatozoa are ideal candidates for whole-cell MS, as they are highly accessible, easily purified and relatively biochemically simple compared to other eukaryotic cells (Dorus et al. 2006; Karr 2007; Oliva, de Mateo, and Estanyol 2009). In addition, MS of sperm cells has proven to be more reliable than other inference methods, such as the use of testis expression as a proxy, for determining which proteins are incorporated into mature spermatozoa. For example, in the *Drosophila melanogaster* sperm proteome (DmSP) almost a quarter of genes were expressed <0.5-fold in the testis compared to the remainder of the fly, meaning that they had half the expression in the testis than in non-reproductive tissues; a reminder that not all DmSP genes are sperm-specific (Wasbrough et al. 2010). The analysis of spermatozoa by MS has now been carried out on a variety of species, including *Caenorhabditis elegans* (Chu et al. 2006), bull (*Bos taurus*) (Peddinti et al. 2008) and rice pollen (*Oryza sativa*) (Dai et al. 2006), and has significantly increased the number of known sperm proteins in a relatively short period of time (Karr 2007; Oliva, De Mateo, and Estanyol 2009; Findlay and Swanson 2010). The current DmSP contains >1,000 proteins, however, prior to MS analysis, only 16 proteins had been experimentally shown to be in *D. melanogaster* sperm, of which only 5 were empirically identified as encoding sperm components (Dorus et al. 2006; Wasbrough et al. 2010).

By providing detailed information about the protein complement of sperm MS has increased our understanding not only of sperm protein composition but also of sperm protein evolution and functional composition (Karr 2007; Oliva, de Mateo, and Estanyol 2009; Findlay and Swanson 2010). Proteomic analyses of the DmSP demonstrate that a large number of genes, with predicted functions, are involved in metabolism and sperm movement (Dorus et al. 2006; Wasbrough et al. 2010). In addition, an enrichment of genes involved in proteolysis is also observed (Dorus et al. 2006; Wasbrough et al. 2010), and this included the sperm leucyl aminopeptidase (S-LAP) gene family. S-LAPs are an extended gene family constituting the most abundant proteins by mass in the DmSP, yet their function in sperm remains unclear (Dorus, Wilkin, and Karr 2011). In general, peptidases have been found in a variety of male and female reproductive tissues (Swanson et al. 2001; Swanson et al. 2004; Prokupek et al.

2009). Many of the genes empirically identified in the DmSP have experimentally documented effects on sperm motility, development and function (Wasbrough et al. 2010), including the well-characterised sperm structural proteins such as βTub85D, βTub56D and αTub84B (Fuller et al. 1988; Hoyle and Raff 1990; Kimble, Dettman, and Raff 1990; Kaltschmidt et al. 1991), and the Y-linked fertility factors, kl-3 and kl-5 (Goldstein, Hardy, and Lindsley 1982; Gepner and Hays 1993; Carvalho, Lazzaro, and Clark 2000) that encode dynein heavy chains essential to the axoneme outer arm structure (Goldstein, Hardy, and Lindsley 1982). MS analysis of the DmSP indicates that the evolution of the sperm proteome is quite distinct from the evolution of other reproductive proteins. For example, while accessory gland proteins (ACP) and other reproductive proteins are rapidly evolving, the DmSP is evolving conservatively and often under purifying selection (Dorus et al. 2006). Furthermore, genes that encode proteins with a structural function or are involved in metabolic pathways, were those most conserved, probably due to pleiotropic constraints (Dorus et al. 2006).

While the Drosophila genus remains the most intensively studied in regards to insect sperm, the proteins involved in fertilization still remain unclear. In mammals, the mouse (*Mus musculus*) represents the best-studied species in terms of sperm protein function. MS analysis has been performed on whole sperm (Baker et al. 2008; Dorus et al. 2010), fractionated sperm membranes (Asano et al. 2010), flagellum accessory structures (Cao, Gerton, and Moss 2006) and regions that may be involved in mediating sperm-egg interactions (Stein et al. 2006). A large proportion of proteins in the *M. musculus* sperm proteome (MmSP) have functions in metabolism and energetics, inter- and intra-cellular transport and catalysis/proteolysis (Baker et al. 2008; Dorus et al. 2010). Analyses of the MmSP have identified proteins with functions in flagellum structure, such as Outer Dense Fibres (*ODF1 and 2*) and multiple tektins (Cao, Gerton, and Moss 2006), as well as a diversity of immune-related proteins (Dorus, Skerget, and Karr 2012), and there are > 50 proteins with reported reproductive phenotypes in males (Dorus et al. 2010). Targeted molecular studies in mouse have identified many of the sperm proteins involved in sperm-egg interactions, including several members of the A Disintegrin And Metalloproteinase (ADAM) family, and several zona pellucida binding proteins (*Zp3r, Zpbp, Zan*) (reviewed in Nixon, Aitken, and McLaughlin 2007). Finally, similar to the DmSP, the MmSP genes functioning in energetic and structural roles are evolving more conservatively than other sperm genes, while genes encoding sperm membrane proteins are generally evolving much more rapidly (Dorus et al. 2010).

While proteomic analysis of individual protein sets (proteomes) is useful, comparative proteomics has the potential to further our understanding of protein composition and protein evolution. Comparative proteomics refers, at its most basic level, to the comparison of different

sets of proteins (i.e. two different proteomes), whether these proteomes are from different cell types or stages, of diseased versus healthy tissue, samples from a series of time points or from different species. There are a limitless number of combinations in which comparative analyses of protein sets could, and have, been used to gain insights that would not be possible from a simple single catalogue of proteins, and include studies that have important implications for human health and food production. For example comparative proteomics have been employed to investigate how the seeds of transgenic rice strains differ from their wild type stains in an attempt to assess the impact of these differences on human health (Xue et al. 2012), to isolate proteins changes that may be important in diseases such as polycystic kidney disease by comparing healthy and diseased tissues (Li et al. 2012), or to classify disease stages, such as in ovarian cancer (Kim et al. 2008), as well as studies into food production including how proteins in eggs differ when stored at different temperatures (Qiu et al. 2012). Initial comparisons of the DmSP and MmSP demonstrated that there is extensive homology between sperm proteins of these species (Dorus et al. 2006; Wasbrough et al. 2010). However, a detailed comparative analysis of these two sperm proteomes has yet to be conducted. Due to its use as a model for human infertility the MmSP has been subjected to targeted molecular studies that have elucidated the roles of specific sperm proteins. Despite *D. melanogaster* having a well-curated genome and the extensive availability of mutant strains, the proteins in the DmSP have not been extensively subjected to this level of investigation. However, *D. melanogaster* is especially amenable to studies, and has reduced ethical, and monetary implications of its use compared to mouse. Therefore, if extensive parallels between the sperm proteome of the current model for human infertility (mouse) and *D. melanogaster* exist, it is possible that *D. melanogaster* may have the potential to be an initial study organism for proteins, and pathways, involved in reproduction generally, but also to identify gene candidates for more targeted studies in mammalian systems. *D. melanogaster* is instrumental in the investigation of many mammalian process (Karr 2007), including cancer (reviewed in Miles, Dyson, and Walker 2011; Rudrapatna, Cagan, and Das 2012), wound healing (reviewed in Razzell, Wood, and Martin 2011) and Alzheimer's disease (reviewed in Mhatre et al. 2013). Here, we perform a comparative analysis of the *D. melanogaster* and mouse sperm proteomes to determine whether there is an overall similarity in functional composition between the spermatozoa of these species, as well as to demonstrate how a comparative approach can provide additional insights compared to analysis of an isolated proteome.

Mouse and Drosophila sperm differ in several key aspects including development, morphology and physiology. Mammalian spermatozoa tend to be relatively small and highly motile cells that can be morphologically divided into three distinct sections (head, midpiece, flagella). The mammalian midpiece contains large numbers of individual mitochondria, while glycolytic

enzymes associated with the fibrous sheath span the length of the flagella (Krisfalusi et al. 2006). Mammalian sperm motility is correlated with competitive ability, and although it seems that glycolysis rather than oxidative-reduction is the main provider of energy (Nascimento et al. 2008), sperm competitive ability is correlated with midpiece volume but not flagella length or volume (Anderson, Nyholt, and Dixson 2005). Conversely, within the Drosophila genus, sperm length, rather than sperm motility, is indicative of sperm competitive ability (Miller and Pitnick 2002), and sperm gigantism is prevalent throughout the genus (Karr and Pitnick 1996). Drosophila sperm length is thought to have co-evolved with female reproductive tract architecture (Miller and Pitnick 2002; Pitnick, Markow, and Spicer 1999). In addition, while the mammalian midpiece contains large numbers of individual mitochondria, Drosophila sperm contain a pair of mitochondrial derivatives, termed the nebenkern, which span the length of the flagellum. While the contribution of this organelle to energy production is unclear it has an important role in the elongation of the sperm tail (Noguchi, Koizumi, and Hayashi 2011). Finally, while mammalian spermatozoa are briefly stored by attachment to the oviductal epithelium before undergoing hypermotility and capacitation (Holt and Fazeli 2010), Drosophila sperm can be stored in specialised storage organs (spermatheca and seminal receptacle) within the female for several weeks (Pitnick, Markow, and Spicer 1999). As such, despite initial similarities (Dorus et al. 2006; Dorus et al. 2010; Wasbrough et al. 2010), a detailed comparative analysis, of their proteome complements could yield variances that reflect these differences in biology.

Despite their morphological differences, sperm have the same function. In initial studies, striking conservation has been observed between mammalian and Drosophila sperm proteomes. Comparisons between the MmSP and the first characterisation of the DmSP observed that >40% of mouse axoneme structure sperm proteins have homology to proteins in the DmSP (Dorus et al. 2006). Similarly, comparisons between the mammalian sperm proteomes and an expanded DmSP found that >65% of DmSP genes possessed a mouse and/or human ortholog, of which 20% were also identified within the respective mammalian sperm proteome, with the clearest conservation within metabolic and energetic pathways (Wasbrough et al. 2010). Both taxa require similar proteins to produce energy, whether for sperm motility or to prolong sperm viability, as well as similar proteins for structuring and movement of the flagella; so differences may be in the amount and use of such proteins. It therefore seems likely that a detailed comparison of proteins that contribute to the sperm proteome of these two species will identify significant similarities between the sperm proteomes. Here we present a detailed comparative analysis of the protein composition of the MmSP and DmSP. We determined that there is general functional conservation between the two species, with significant similarity in the proportion of genes allocated to specific functions. Further, we identified similar proteins

involved in immunity and proteolysis in both proteomes, which may indicate parallel evolution beyond simple functional requirements. Finally, we identified a significant under-representation of X-linked sperm proteome genes in both species, but observed that the primary cause of this differed between the two genomes.

2.2 Materials and methods

2.2.1 Proteomic datasets

The protein composition of *D. melanogaster* and *M. musculus* spermatozoa have previously been analysed using mass spectroscopy (MS) (Cao, Gerton, and Moss 2006; Dorus et al. 2006 Stein et al. 2006; Baker et al. 2008; Dorus et al. 2010; Wasbrough et al. 2010). To obtain the initial *D. melanogaster* sperm proteome (termed DmSP-I) mature spermatozoa were purified from the seminal vesicles of 25-50 virgin males, and subjected to MS, resulting in the identification of 342 unique proteins (Dorus et al. 2006). The DmSP was subsequently expanded due to improvements in MS technology and refinement of experimental protocol yielding the DmSP-II (Wasbrough et al. 2010). The DmSP-II dataset was generated from sperm isolated from the seminal vesicles of 75 virgin males. For processing, these samples were centrifuged and washed to produce a pellet. Pellets were re-suspended in sample buffer containing 1× Nupage® reducing agent (Invitrogen, Inc) and the protein quantity of the suspension was determined. Fifty micrograms (50µg) of protein was separated on a 1-Dimensional SDS-PAGE gel. After electrophoresis, the gel lane was divided once vertically, then horizontally divided into 16 segments; each segment was independently subjected to MS. The DmSP-II was analysed using both Sequest and X! Tandem software and the results reconciled using Scaffold, while the DmSP-I was analysed by Sequest (Dorus et al. 2006; Wasbrough et al. 2010). Both Sequest and X! Tandem generate *in silico* datasets of potential MS spectra, by determining all digested peptide fragments potentially being produced by a genome and therefore all potential spectra. It is these spectra that are used to identify the peptides observed by the MS. The DmSP-II identified 956 sperm proteins, of which 189 (out of 342) were also identified in the DmSP-I, and network analyses demonstrated that while a small number of new functional categories were identified, the majority of the DmSP-II represented an expansion of the original DmSP-I (Wasbrough et al. 2010). Therefore, for our analyses, we used the 1,108 proteins that have been identified in the DmSP-I and/or the DmSP-II (from this point forward termed the DmSP). For the mouse sperm proteome (MmSP) we used the proteome database collated by Dorus *et al.* (2010) which is comprised of MS data from two whole spermatozoa analyses (Baker et al. 2008; Dorus et al. 2010), as well as data from two further MS studies that were limited to the flagella accessory structure (Cao, Gerton, and Moss 2006) and fractions of the cell involved in interactions between sperm and egg; including acrosome and cell membrane (Stein et al. 2006). Together, these studies have empirically identified 996 unique proteins in MmSP.

2.2.2 Functional Composition Analysis

The functional composition comparison was based on the proportion of the sperm proteome in each biological processes (BP) gene ontology (GO) category in PANTHER (http://www.pantherdb.org/). The numbers of unmapped gene identifiers were classified as genes with unknown (unannotated) GO functions. Comparisons between the proportions of sperm proteome genes within each BP GO category were assessed using Fishers' 2-tailed test, with Bonferroni correction. In addition, each sperm proteome was assessed for enrichment of GO categories using GOEAST (http://omicslab.genetics.ac.cn/GOEAST/) using the Bonferroni correction for multiple testing and a minimum significance level of 0.05. Finally, we surveyed the DmSP for genes with curated functions in immune processes. We obtained all GO annotations for DmSP and MmSP genes from FlyBase (www.flybase.org) and MGI (http://www.informatics.jax.org/), respectively. A survey of these genes was conducted for any terms related to, or descended from, the GO term immune system processes (GO: 0002376). We statistically assessed whether the contribution of immune-related genes to the sperm proteome was similar for both the DmSP and MmSP utilising two-tailed χ^2, with Yates' correction, tests.

2.2.3 Domain Composition Analysis

We created a catalogue of annotated protein domains encoded by genes in the DmSP and MmSP, based on IPR classification (http://www.ebi.ac.uk/interpro/) obtained from FlyBase and MGI, respectively. In order to identify genes with potentially parallel functions we undertook a comparison of the sperm proteomes to detect protein domains that were encoded by genes in both proteomes. An additional survey was conducted to classify the variety of protease families and their inhibitors potentially encoded by sperm proteome genes, and to determine the relative importance of each peptidase family. We statistically assessed whether the contribution of peptidases and inhibitors to the sperm proteome was similar for both the DmSP and MmSP utilising two-tailed χ^2, with Yates' correction, tests.

2.2.4 X-linked proteome genes

To assess whether there is an under or over-representation of sperm protein encoding genes on the X chromosome, we determined the expected number of X-linked proteome genes in each species based on the size of the X chromosome compared to the autosomes. Chromosomal locations for all MmSP and DmSP genes were obtained from FlyBase and MGI, respectively. The relative contribution of the X chromosome to the genome was calculated in both species based on the size of the X chromosome relative to the autosomes (not including the Y chromosome or the Drosophila chromosome 4 due to the small number of genes). The proportion of genome consisting of the X chromosome was taken as the proportion of X-linked

sperm proteome genes expected if genome distribution was random. Lengths of chromosomes (bp) were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). Significant differences between the number of observed and expected X-linked sperm proteome genes were assessed using two-tailed χ^2 test, with Yates' correction.

2.2.5 Expression of X-linked sperm proteome genes during spermatogenesis

Expression data was available for each stage of mammalian spermatogenesis: mitosis (Type A spermatogonia, Type B spermatogonia), meiosis (Pachytene spermatocytes) and post-meiosis (Round spermatids), from the Mammalian reproductive genetics database (mrg.genetics.washington.edu) using the Affymetrix GeneChip Mouse Genome 430 2.0 Array for 26 of the 29 X-linked MmSP genes. Similarly, expression data was available for samples enriched for specific stages of *D. melanogaster* spermatogenesis from Vibranovski *et al.* (2009), for 150 of the 170 X-linked DmSP genes. Significant differences in the distribution of expression measurements between meiotic stages and each other stage of spermatogenesis was assessed using Kolmogorov-Smirnov tests.

2.3 Results

2.3.1 Comparative analysis of function

Utilising PANTHER (http://www.pantherdb.org/), an online database that classifies gene function using both direct experimental evidence and evolutionary relationships, we demonstrated that the MmSP and DmSP have substantial functional conservation, with both proteomes containing similar proportions of genes within the GO categories provided (Figure 2.1). The majority of MmSP and DmSP genes function in metabolism and the generation of precursor metabolites and energy, with a substantial number of MmSP and DmSP genes classified as having functions in the cell cycle, and developmental processes and cellular organization. However, a significantly larger proportion of DmSP genes were classified as unknown compared to the MmSP ($p < 0.0001$). This proportional difference is similar to the proportion of genes within each genome with uncharacterized functions. In addition, the MmSP contained a significantly higher proportion of genes with functions in immune processes ($p < 0.0001$) and response to stimuli ($p < 0.0001$) compared to the DmSP.

The overall extensive functional similarity between the MmSP and DmSP is supported by a GO enrichment analysis, performed in GOEAST (http://omicslab.genetics.ac.cn/GOEAST/), an online software that identifies significantly enriched GO categories within gene sets. Both proteomes were significantly enriched for genes involved in metabolic processes, including both oxidative phosphorylation and glycolysis. They were also enriched for genes involved in cellular developmental processes, cell differentiation and cellular component organization. The DmSP, but not the MmSP, was enriched for genes involved in microtubule cytoskeleton organization and phosphorylation. However, the MmSP is enriched for genes involved in several categories in which the DmSP is not, including: spermatogenesis, sperm-egg recognition, binding of sperm to the zona pellucida and fertilization.

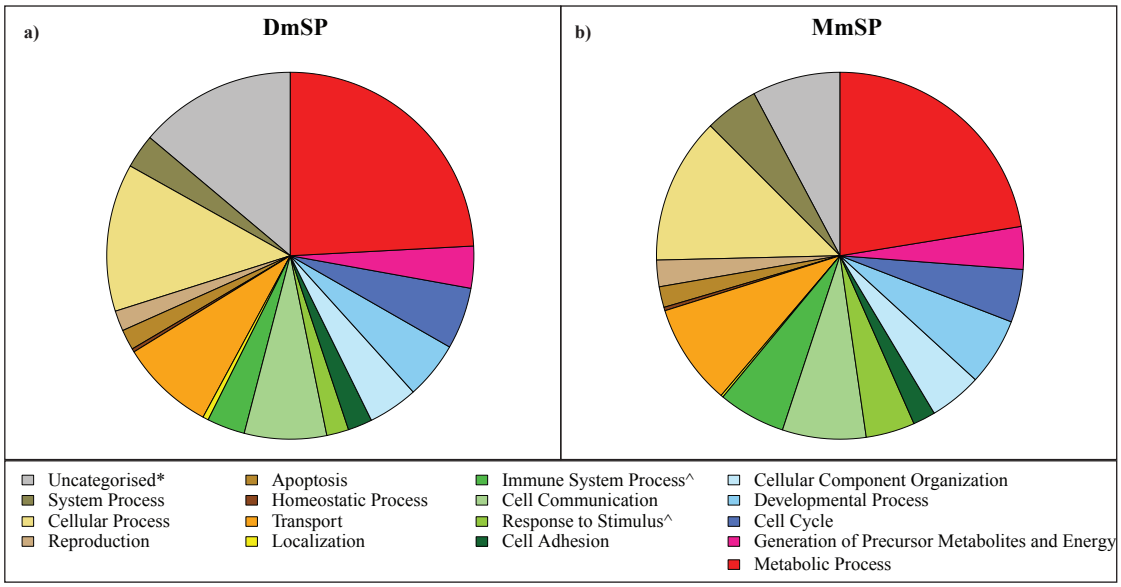

Figure 2.1 Functional comparison between the Drosophila and mouse sperm proteomes.

The number of *Drosophila melanogaster* and *Mus musculus* sperm proteome (DmSP and MmSP, respectively) genes within each gene ontology category was obtained from PANTHER (www.pantherbd.org). Genes not annotated in PANTHER were classified as "uncharacterized". Proportions were based on the total number of genes within each sperm proteome, allowing genes to be present in multiple categories. Significant differences between proportions of DmSP and MmSP genes within functional categories were determined using 2-tailed Fishers' exact test, with Bonferroni correction for multiple testing. Gene ontology categories containing a significantly higher (*) or lower (^) proportion of sperm proteome genes in the DmSP compared to the MmSP are indicated in the legend.

2.3.2 Immunity proteins of the sperm proteome

A survey of DmSP genes identified 20 genes with a variety of GO categories related to immune functions (Table 2, Appendix II). Including two (*Hel89* and *αTub84*) that have been implicated in sperm function based on the identification of male sterile alleles (www.flybase.org) and several (*Toll-4, Tollo, spz3*) that are in the Toll-pathway. The presence of genes with immunity functions in mammalian spermatozoa have been reviewed by Dorus *et al*. (2012), however, following the same procedure as for genes within the DmSP, we conservatively identified 63 MmSP genes with annotated GO functions related to immunity (Table 2.1). As predicted by the PANTHER analysis of the two proteomes, the MmSP has a significantly higher proportion of genes with immune related functions compared to the DmSP ($\chi^2 = 27.103$, $p < 0.0001$). Of the MmSP genes, 10 have reported phenotypes associated with reproduction. Of particular interest are *Gpx4* and *Camk4*, which have phenotypes affecting sperm head, flagellum, mitochondrial sheath and principle piece morphology, and *Hsp90aa1*, which has phenotypes affecting sperm

production (http://www.informatics.jax.org/). In addition, similar to the DmSP, the MmSP also contains three genes (*Hspd1, Pdpk1, Arf6*) that are in the Toll pathway. Finally, the MmSP contains a number of genes involved in the regulation of NF-kappaB import into the nucleus (*Prdx1, Ptgs2*), binding (*Hspa1b, Psma6*) and transcriptional activity (*Clu, Itgb2, Psma6, Prdx2*), and regulation of I-kappaB kinase/NF-kappaB cascade (*Eef1d, Vapa, Gstp1*). NF-kappaB is important in the transcription of DNA in response to stimuli including infection.

Table 2.1 MmSP genes with immune functions

Gene Symbol	GO Accession	GO Term
Immune Functions		
Was, Ighg1	06955, 50778	Immune response
C3, Cd55, Hc, Apoa4*	45087, 02227	Innate immune response
Hspa1b, Ppp3r2, Ighg1 (bacterium), Cotl1(fungus)	06952, 42742, 50832	Defense response
Hpx, Ighg1	02925, 19731, 02455	Humoral immune response
C3, Camk4*, Hc, Hist1h2ba, Park7, Psma1, Psmb4, Ptgs2*, Ace*, Fabp4, Hspd1, Gstp1, Agt, Gpx4**	06954, 02639, 02674, 02862 50727	Inflammatory response
Phb, Pde1b*	71354, 36006	Cellular response
Response to Bacterial molecules		
B2m	02237	Response to molecule of bacterial origin
Ace, Prdx2, Ptgs2**	32496	Response to lipopolysaccharide
Toll Pathway		
Hspd1, Pdpk1, Arf6*	02755, 34122, 34143	Toll-like receptor signaling pathway
Adaptive immunity		
Ighg1, Ppp2r1a, Ap3d1, Psme1, B2m, Bat3*	03823, 19882, 19884, 02481, 02474, 48007	Antigen activity
C3, Hc, Cd46*, Cd55, Ighg1, Krt1, C1qbp, Cfh*	06956, 06957, 06958, 01867, 01848, 01849, 01851, 45916, 30449	Complement activity
*Drd2**	34776	Response to histamine
Ighg1, Hpx	42571, 16064, 02639	Immunoglobulin
Hspd1	48291	Isotype switching to IgG isotypes
Hspd1	32727, 32729	Interferon production
Gstp1, Apoa1, Gc, Hspd1, Apoa2, Timm50	32691, 50713, 32715, 32733, 32735, 32755, 45416, 19976, 05134, 19982	Interleukin (regulation of production and secretion; binding)
Calr, Atp5a1, Atp5b, B2m	42824, 42288, 42612, 02502	MHC
Apoa1, Apoa2	02740	Regulation of Cytokine secretion
Gstp1	71638	Negative regulation of monocyte chemotactic protein-1 production

Table 2.2 Continued

Immune Cells		
Hspd1, Gc	42113, 02368, 42100, 45579	B-Cell activity (inc. differentiation)
Apoa4, Itga5, Itgb2, Hc, Lta4h, Gstp1, Nme2, Tex101	07159, 02523, 19370, 06691, 04463, 70664, 02762, 02696	Leukocyte activity (including proliferation, differentiation, activation)
Hprt	46651	Lymphocyte proliferation
Hspd1, Hc	43032, 10760	Macrophage activation and regulation of chemotaxis
Pdpk1, Ywhaz*	43304, 02553	Regulation of mast cell degranulation and histamine secretion
*Itgb2, Prdx1, Ap3d1**	30101, 42267, 51138	Natural Killer Cell activity
Ace, Gstp1, Itgb2*	02446, 2000429, 30593	Neutrophil mediated immunity, aggregation and chemotaxis
Itgb2, Myh9, Mink1 Prdx2, M6pr, Gc, Hsp90aa1, Hspd1, Was, B2m, Hspa1b*	50798, 01768, 45581, 42104, 32831, 45585, 45060 50870, 42110, 01916, 42098	T cell activity (including proliferation, differentiation, activation)
AntiViral functions		
Hsp90b1	46790	Virion binding
Hdac1, Pfn1	43922, 50434	Regulation of viral transcription
Srpk2 *Ppia*	45071, 45070, 45069	Regulation of viral genome replication
Dynlt1b	19060	Intracellular transport of viral proteins in host cell

* Reproductive associated phenotype (http://www.informatics.jax.org/)

2.3.3 Proteases and protease inhibitors in the sperm proteome

Utilising GOEAST, we found that the MmSP is also enriched for genes involved in peptidase activity ($p = 1.12e-7$) and regulation of peptidase activity ($p = 0.028$). A survey of the DmSP identified 60 proteins with peptidase domains (Table 3, Appendix II), which is statistically indistinguishable from the proportion of genes with peptidase domains in the MmSP (Table 2.2) (DmSP: 5.5%; MmSP: 5.7%; $p = 0.9188$; based on the number of sperm proteome genes with documented domain information). In both proteomes the majority of peptidases contain metallopeptidase or serine peptidase domains. Similarly, in both proteomes, the majority of protease inhibitors contain domains that inhibit serine proteases.

Table 2.2 MmSP genes with peptidase or proteinase inhibitor domains

Peptidase Domain	Gene Symbols
Metallopeptidases	
Peptidase M1, M2	*Lta4h, Ace*
Peptidase M12B	*Adam1b, Adam2, Adam3, Adam4, Adam5, Adam6a, Adam6b Adam24*
Peptidase M13, M14, M16 (or M16C), M17, M18, M19	*Mmel1, Cpa5, Pitrm1, Uqcrc1. Uqcrc2, Lap3, Dnpep, Dpep3*
Peptidase M20, M24, M28	*Pgcp, Xpnpep1*
Peptidase M42, M49	*Dnpep, Dpp3*
Serine peptidases	
Chemotrypsin (S1A and S1/S6; S1C)	*4930519F16Rik, Acr, Htra2, Pig, Prss21, Prss32, Prss46, Prss50, Prss52, Prss54, Try10, Htra2*
Non-chemotrypsin (S8, S8A S8/S53, S10, S28, S37, S60	*Cpul, Prcp, Ltf, Trf, Pcsk6, Tpp2*
Other peptidases	
Peptidase T1A, T2	*Psmb5, Psmb6, Psmb7, Asrql1*
Peptidse C12	*Uchl1, Uchl3, Uchl3, Uchl5*
Peptidase C19	*Usp7, Usp4*
Peptidase C48	*Senp8*
Proteinase inhibitor	
Leucine-rich repeat, ribonuclease inhibitor subtype	*Lrrc34*
Protease inhibitor I8, cysteine-rich trypsin inhibitor-like*	*Muc5b, Zan*
Proteinase Inhibitor I1, Kazal*	*Spink5*
Proteinase inhibitor, carboxypeptidase propeptide	*Cpa5*
Serpin domain	*Agt, Gm46, Serpina1a, Serpina1b, Serpina1d, Serpina1e, Serpina1f, Sperpina3k, Serpina5, Sperpinb6a*

* Inhibits S1, S8, M4 family of proteases

2.3.4 Under-representation of X-linked sperm proteome genes

Meiotic sex chromosome inactivation (MSCI), in which the sex chromosomes are transcriptionally silenced during meiosis, is a well-established phenomenon in mouse, and is believed to influence the lack of X-linked male-biased genes, however, its presence in Drosophila is debated (Hense, Baines, and Parsch 2007; Vibranovski et al. 2009; Meiklejohn et al. 2011; Mikhaylova and Nurminsky 2011). To determine whether an under-representation of X-linked genes that encode sperm proteins (sperm genes) exists, we assessed whether the expected number of X-linked sperm genes significantly differs from the observed number based on the proportionate size of the X chromosome. In mouse, the X chromosome composes ~6% of the mouse genome, based on both its relative size and gene contribution, as such using the number of MmSP genes with known chromosomal locations (986/996) 62 are expected to reside on the X chromosome, which is significantly higher than the number of observed X-linked MmSP genes (29/986; $\chi^2 = 11.797$, $p = 0.0006$). However, the Drosophila X chromosome comprises ~19% or ~16% of the Drosophila genome based on its relative size or gene content, respectively. Utilising the number of DmSP genes with known chromosomal locations (1108/1108), 209 are expected to reside on the X chromosome based on the size of the X chromosome. This is significantly higher than the number of observed X-linked DmSP genes (170/1108; $\chi^2 = 4.596$, $p = 0.0320$). However, 173 are expected to reside on the X chromosome based on the proportion of X-linked versus autosomal genes. This is not significantly different from the observed number of X-linked DmSP genes ($\chi^2 = 0.014$, $p = 0.9065$).

2.3.5 Analysis of X-linked sperm proteome gene expression during spermatogenesis

We observed that the majority of X-linked MmSP genes had their lowest recorded expression during meiosis (15 out of 26) and that none had their maximum-recorded expression during this stage. However, in Drosophila, the stage of spermatogenesis in which the lowest expression was recorded was relatively equally distributed between mitosis (n = 45), meiosis (n = 52) and post-meiosis (n = 53) with the majority of X-linked DmSP genes, had their maximum-recorded expression during mitosis and post-meiosis (123 out of 150). In mouse, average expression during meiosis was significantly lower than the average expression during mitosis ($D = 0.4231$, $p = 0.013$) and post-meiosis ($D = 0.4615$, $p = 0.005$) (Figure 2.2a). Conversely, in Drosophila, average expression during meiosis was not significantly lower than average expression during mitosis ($D = 0.1267$, $p = 0.167$) or post-meiosis ($D = 0.0733$, $p = 0.800$) (Figure 2.2b).

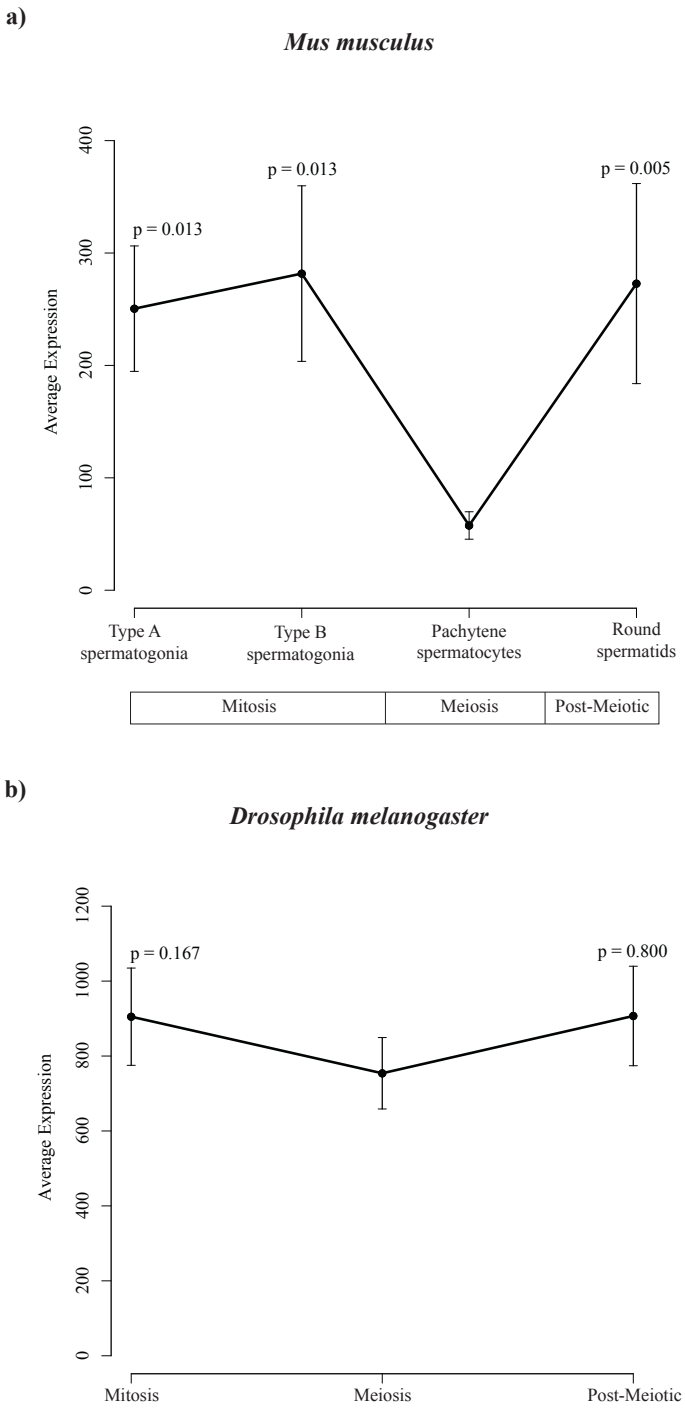

Figure 2.2 Average expression of X-linked sperm proteome genes throughout spermatogenesis. Average expression (± standard error) for each stage of spermatogenesis for (a) X-linked MmSP genes and (b) X-linked DmSP genes. Expression data was obtained for each stage of spermatogenesis from the Mammalian reproductive genetics database (mrg.genetics.washington.edu) and Vibranovski *et al.* (2009) for mouse and Drosophila, respectively. P-values given indicate the difference between mitosis or post-meiosis, and meiosis based on Kolmogorov-Smirnov tests.

2.4 Discussion

Spermatozoa are an incredibly specialised cell type functioning to deliver the paternal genetic contribution in sexually reproducing organisms and are considered one of the most morphologically diverse cell types (Pitnick, Hosken, and Birkhead 2009). Due to their important role in sexual reproduction sperm are under intense sexual selection, which is likely to have a significant effect on their protein composition and evolution. MS analysis of sperm has provided detailed information of their protein composition, function and evolution (Karr 2007; Oliva, De Mateo, and Estanyol 2009; Findlay and Swanson 2010) however, a detailed comparative proteomic analysis of sperm had not been conducted. Initial analyses determined that despite differences in morphology and physiology there was significant homology between proteins found in the DmSP and MmSP (Dorus et al. 2006; Wasbrough et al. 2010). Consistent with these studies we observed that there is extensive functional conservation of the proteins that contribute to mature spermatozoa, as well as cases of parallel evolution, as demonstrated by the abundance of genes with putative functions in immunity and proteolysis in both proteomes. In addition, because detailed studies performed in mice demonstrate the importance of many of these genes in sperm-oocyte interactions, we identified several candidate genes for these interactions in Drosophila (*TepII, TepIV, Toll, spz, Ppn, stl, tace, spn47C* and *spn43Ab*), suggesting that a detailed molecular analysis of these genes may provide important information on sperm-egg interactions; which in turn may provide further information on the proteins involved in mammalian fertility and reproduction. In the remaining sections we will lay out the extensive parallels that we have observed between the mouse and Drosophila sperm proteomes, including the overall functional composition and the existence of sperm-specific immune and proteolytic genes.

2.4.1 Functional conservation of the sperm proteome composition

Initial analyses of the DmSP and MmSP demonstrated that there is evolutionary constraint in sperm proteome evolution and that there is extensive homology among the sperm proteins of these two species (Dorus et al. 2006; Dorus et al. 2010; Wasbrough et al. 2010). These results are consistent with our comparative analysis of the mouse and Drosophila sperm proteomes. Comparison of the functional composition of the MmSP and DmSP demonstrated that similar proportions of genes were involved in the same biological functional processes (Figure 2.1) with the exception of a higher proportion of unknown genes and a lower proportion of genes with functions in immunity and response to stimuli in the DmSP, compared with MmSP. Genes involved in metabolism and the generation of precursor metabolites and energy represent the largest GO category in both Drosophila and mouse, potentially highlighting the high-energy demands of this cell type. It is also noteworthy that in analysis of the DmSP, homology with mammalian proteins was greatest for genes involved in metabolic processes (Wasbrough et al.

2010). Further, a substantial number of MmSP and DmSP genes were classified as having functions in the cell cycle, development processes and cellular organisation. This is not surprising given that spermatozoa undergo dramatic morphological changes during spermatogenesis. Variations however, in the enrichment of genes involved in microtubule cytoskeleton organization between DmSP and MmSP were observed with significant enrichment in DmSP, but not the MmSP. This likely reflects differences in sperm biology, as the association between microtubules and the nebenkern is essential for correct flagellum development during Drosophila spermatogenesis (Noguchi, Koizumi, and Hayashi 2011).

The largest difference in functional composition of the sperm proteomes is from differences in the proportion of sperm genes with functions in immunity and response to stimuli. However, it should be noted that there is extensive overlap between the MmSP genes classified as immunity and stimuli genes. The difference between DmSP and MmSP may therefore be due to (1) differences in the two proteomes reflecting differences in the sperm biology of the two species, (2) more intensive study of the MmSP due to its relevance to human infertility and other diseases, and (3) the presence of adaptive immunity in mouse but not in Drosophila. Consistent with previous reports, adaptive immunity proteins extensively contributed to the total number of immunity proteins in the MmSP (Dorus, Skerget, and Karr 2012). In addition, the enrichment analysis reported that the MmSP, but not the DmSP, is enriched for genes with functions in spermatogenesis, sperm-egg recognition, binding to the zona pellucida and fertilisation. Again, this may be due to differences in underlying sperm biology, but more likely represents differences in the intensity to which these processes have been studied in mammals and insects.

2.4.2 Diversity of immune-related genes in both mouse and Drosophila sperm proteomes

The trade-off between reproduction and immunity, and the post-mating immune response of females, have been intensively studied in Drosophila (Lawniczak and Begun 2004; Mcgraw et al. 2004; Fedorka et al. 2007; Lawniczak et al. 2007; Kapelnikov et al. 2008; Innocenti and Morrow 2009). Meanwhile, a diverse array of immune genes have been identified with sperm-specific functions in mammals (reviewed in Dorus, Skerget, and Karr 2012). We have consistently observed that the MmSP contains a number of immune-related genes that have phenotypes associated with reproduction (Table 2.1). While the MmSP has a higher proportion of genes with functions in immunity, several of the DmSP genes have similarities either in function or protein domains encoded to genes in the MmSP. Therefore, we propose that these DmSP genes may have similar sperm-specific functions to their MmSP counterparts, and may be excellent candidates for more targeted studies to elucidate proteins involved in insect fertilisation.

In mammals, there has been recent interest in the sperm-specific functions of complement regulators (Harris, Mizuno, and Morgan 2006; Dorus et al. 2010; Dorus, Skerget, and Karr 2012). Two of these regulators, *Cd55* and *Cd46*, have been identified in the MmSP and both are localised to the acrosomal region of spermatozoa. As such, although *Cd55* also has weak expression across the sperm surface, it is unlikely that these complement factors function in defence against either female-mediated complement or pathogens (Inoue et al. 2003; Harris, Mizuno, and Morgan 2006; Mizuno et al. 2007). *Cd46* has a documented sperm-specific function in the control of the acrosome reaction (Clift et al. 2009). Surprisingly, given that insects do not have complement-mediated immunity, a family of 6 proteins have been identified in Drosophila with homology to thioester-containing proteins in mammals (Blandin and Levashina 2004). Two of these genes (*TepII* and *TepIV*) were identified in the DmSP. Thioester-containing proteins are import components of the complement-mediated adaptive immune system in vertebrates (Blandin and Levashina 2004). In *D. melanogaster*, *TepII* and *TepIV* have annotated roles in the antibacterial humoral response, are up-regulated during bacterial immune challenge (Blandin and Levashina 2004) and, although a reproductive function has not yet been identified for these genes, up-regulation is detected in the female after mating (Innocenti and Morrow 2009). Given their homology to complement proteins in mammals, the number of complement associated proteins in the MmSP, and the documented sperm-specific roles of certain mammalian complement regulators, we propose that *TepII* and *TepIV* may have, an as yet, undocumented, important role in Drosophila sperm and reproduction in general.

Both the MmSP and DmSP contain several proteins with functions in, or relating to, the Toll signalling pathway. The Toll-pathway is one of two signalling pathways in the innate immune system. The activation of the Toll pathway occurs after the cleavage of *spätzle* (*spz*) from the Toll receptor located in the cell membrane via a serine protease cascade, resulting in a further series of intra-cellular protein interactions and culminating in the activation of NF-kappa transcription factors and the production of AMPs (Leclerc and Reichhart 2004). It is therefore noteworthy that regulators of proteolysis, including serine proteases, are found in both the MmSP and DmSP (discussed below), and that the Toll-signalling pathway has been previously implicated in the regulation of a reproduction related protein. In Drosophila, *sex peptide*, an ACP that affects the female post-copulatory response (Chapman et al. 2003), is dependent upon the Toll-pathway (Peng, Zipperlen, and Kubli 2005) and therefore in turn on serine protease activity. Together, these observations suggest that the Toll pathway may have a significant role in reproduction, and those sperm genes with documented functions within this pathway are ideal candidates for initial studies into this role.

2.4.3 Regulators of proteolysis are found in both MmSP and DmSP

Proteolysis is the breakdown of proteins into smaller peptides and has a role in many biological processes including signalling pathways such as the Toll-pathway. Regulators of proteolysis are likely to be important in reproductive biology as they are abundant in the reproductive systems of both males and females of a variety of taxa (Swanson et al. 2001; Swanson et al. 2004; Prokupek et al. 2009) and are known to regulate reproductive proteins, such as *sex peptide* (Peng, Zipperlen, and Kubli 2005). Proteins involved in male-female interactions within the reproductive tract are likely co-evolving. Consistent with this is the observation that serine protease inhibitors are present in the DmSP (Appendix II, Table 3) and are highly transcribed in the Drosophila female sperm storage organs (Prokupek et al. 2009). Further, there is the potential that such interacting proteins are evolving under sexually antagonistic conditions. One prediction of antagonistic co-evolution is that male and female reproductive proteins may be functionally similar but males and females produce distinct sets of proteins. This allows different genes to be under different selection optima (male versus female biased) while still interacting within the same processes. Consistently, the serine protease inhibitors in the DmSP are not identified among the highly transcribed genes within the female sperm storage organs. Finally, the MmSP and DmSP contain similar proportions and types of proteases and protease inhibitors, suggesting that similar regulation of interacting reproductive proteins occurs in both species.

Both of the studied sperm proteomes contain large numbers metallopeptidases. In the DmSP this is largely attributable to the presence of the S-LAP gene family, while in the MmSP this is largely due to the presence of members of the ADAM family of proteins. The S-LAPs are the most abundant proteins by mass in the DmSP (Dorus et al. 2006) and all seven members are testis-specific (Dorus, Wilkin, and Karr 2011). Each S-Lap contains a catalytic domain, however, it is within this domain that the majority of the sequence divergence within this family has occurred. Despite this loss of enzyme activity these genes have been retained and duplicated over evolutionary time suggesting that a new, and important sperm-specific function has evolved (Dorus, Wilkin, and Karr 2011). In the case of the S-LAP family, there are parallels to the mammalian crystallin family of lens proteins that lost enzymatic activity but acquired a new structural role. Therefore, it has been suggested that the S-LAP may have been selectively retained due to a newly acquired structural function which may have been associated with changes in sperm morphology during the radiation of this taxa (Dorus, Wilkin, and Karr 2011). Conversely the ADAM family have documented functions in sperm-egg recognition and binding (reviewed in Klein and Bischoff 2011). While it is improbable that S-LAP and ADAM proteins have similar functions in sperm, we do observe several DmSP genes with domain

similarities to Adam family members. These genes (*Ppn, stl* and *tace*) contain Peptidase M12B domain ADAM/reprolysin domain (IPR001590), which is found in all members of the Adam family of proteins. We propose that these genes are candidates for studies of the proteins involved in sperm-egg interactions, a topic where currently, very little is known.

2.4.4 The X-linked sperm proteome genes

In both mammals and Drosophila there is a documented under-representation of testis or male expressed genes on the X chromosome compared to the autosomes (Betrán, Thornton, and Long 2002; Parisi et al. 2003; Khil et al. 2004; Sturgill et al. 2007). Consistent with these studies, we observed that in mammals there are significantly fewer X-linked genes encoding sperm proteins than expected, based on the relative sizes of the X chromosome and autosomes. However, in Drosophila an under-representation is only observed based on the size of the chromosome but not on the relative gene content of the X chromosome compared to the autosomes. In mouse, studies have shown that the under-representation on the X chromosome is not observed in genes expressed either pre- or post-meiotically (Khil et al. 2004; Mueller et al. 2008). This has lead to the suggestion that meiotic sex chromosome inactivation (MSCI), where the sex chromosomes are transcriptionally inactive during meiosis, is the primary driver of this under-representation in mouse; primarily because the X chromosome is an unfavourable location for any gene required during meiosis. Consistent with this hypothesis, we observed that X-linked MmSP genes have a lower average expression during meiosis than during either mitosis or post-meiotic stages of spermatogenesis. However, we do not observe a similar reduction during meiosis for Drosophila X-linked sperm genes. Therefore, it does not seem likely that MSCI, if it occurs in Drosophila (Hense, Baines, and Parsch 2007; Vibranovski et al. 2009; Meiklejohn et al. 2011; Mikhaylova and Nurminsky 2011), is the cause of the under-representation of X-linked sperm genes in Drosophila.

2.4.5 Summary

Spermatozoa are a unique cell type that is under intense sexual selection due to its important role in reproduction. Mass spectroscopy has become an invaluable tool in the study of sperm protein composition and has provided detailed catalogues of proteins empirically identified in the mature spermatozoa of many species; allowing both the study of sperm composition and the evolution of sperm proteins (Oliva, De Mateo, and Estanyol 2009). Here, we provide a detailed comparative analysis of the *M. musculus* and *D. melanogaster* sperm proteomes. We observed that consistent with the previously reported homology (Dorus et al. 2006; Wasbrough et al. 2010) there is extensive functional conservation between the sperm proteomes of the two species. In addition, we provide evidence that proteins involved in proteolysis and immunity may have important sperm-specific functions in both species. Lastly, we report that the under-

representation of X-linked sperm genes in mouse, but not Drosophila, is likely to be driven by MSCI. This highlights the fact that although current observations (here the genomic distribution of sperm genes) are similar, the evolutionary causes may be different. Although functional similarities between the proteomes are observed between these distantly related species, it does not imply that these similarities are due to the same drivers of selection in each case. Although this comparison of two distantly related sperm proteomes has proved informative, further studies between closely related species will be necessary to determine whether the trends observed here are common to all sperm proteomes.

2.5 References

Anderson, M. J., J. Nyholt, and A. F. Dixson. 2005. "Sperm Competition and the Evolution of Sperm Midpiece Volume in Mammals." *Journal of Zoology* 267 (2): 135–142.

Asano, Atsushi, Jacquelyn L Nelson, Sheng Zhang, and Alexander J Travis. 2010. "Characterization of the Proteomes Associating with Three Distinct Membrane Raft Sub-types in Murine Sperm." *Proteomics* 10 (19): 3494–3505.

Baker, Mark A, Louise Hetherington, Gabi M Reeves, and R John Aitken. 2008. "The Mouse Sperm Proteome Characterized via IPG Strip Prefractionation and LC-MS/MS Identification." *Proteomics* 8 (8): 1720–1730.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. "Retroposed New Genes Out of the X in Drosophila." *Genome Research* 12 (12): 1854–1859.

Blandin, Stephanie, and Elana A Levashina. 2004. "Thioester-containing Proteins and Insect Immunity." *Molecular Immunology* 40 (12) : 903–908.

Boitrelle, F, F Ferfouri, J M Petit, D Segretain, C Tourain, M Bergere, M Bailly, F Vialard, M Albert, and J Selva. 2011. "Large Human Sperm Vacuoles Observed in Motile Spermatozoa Under High Magnification: Nuclear Thumbprints Linked to Failure of Chromatin Condensation." *Human Reproduction* 26 (7): 1650–1658.

Cao, Wenlei, George L Gerton, and Stuart B Moss. 2006. "Proteomic Profiling of Accessory Structures from the Mouse Sperm Flagellum." *Molecular & Cellular Proteomics* 5 (5): 801–810.

Carvalho, A B, B P Lazzaro, and A G Clark. 2000. "Y Chromosomal Fertility Factors Kl-2 and Kl-3 of Drosophila Melanogaster Encode Dynein Heavy Chain Polypeptides." *Proceedings of the National Academy of Sciences of the United States of America* 97 (24): 13239–13244.

Chapman, Tracey, Jenny Bangham, Giovanna Vinti, Beth Seifried, Oliver Lung, Mariana F Wolfner, Hazel K Smith, and Linda Partridge. 2003. "The Sex Peptide of *Drosophila melanogaster*: Female Post-mating Responses Analyzed by Using RNA Interference." *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9923–9928.

Chu, Diana S, Hongbin Liu, Paola Nix, Tammy F Wu, Edward J Ralston, John R Yates, and Barbara J Meyer. 2006. "Sperm Chromatin Proteomics Identifies Evolutionarily Conserved Fertility Factors." *Nature* 443 (7107): 101–105.

Clift, Leanne E, Petra Andrlikova, Michaela Frolikova, Pavel Stopka, Josef Bryja, Brian F Flanagan, Peter M Johnson, and Katerina Dvorakova-Hortova. 2009. "Absence of Spermatozoal CD46 Protein Expression and Associated Rapid Acrosome Reaction Rate in Striped Field Mice (Apodemus Agrarius)." *Reproductive Biology and Endocrinology* 7: 29.

Dai, Shaojun, Lei Li, Taotao Chen, Kang Chong, Yongbiao Xue, and Tai Wang. 2006. "Proteomic Analyses of Oryza Sativa Mature Pollen Reveal Novel Proteins Associated with Pollen Germination and Tube Growth." *Proteomics* 6 (8): 2504–2529.

Dorus, Steve, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L Karr. 2006. "Genomic and Functional Evolution of the *Drosophila melanogaster* Sperm Proteome." *Nature Genetics* 38 (12): 1440–1445.

Dorus, Steve, Sheri Skerget, and Timothy L Karr. 2012. "Proteomic Discovery of Diverse Immunity Molecules in Mammalian Spermatozoa." *Systems Biology in Reproductive Medicine* 58 (4): 218–228.

Dorus, Steve, Elizabeth R Wasbrough, Jennifer Busby, Elaine C Wilkin, and Timothy L Karr. 2010. "Sperm Proteomics Reveals Intensified Selection on Mouse Sperm Membrane and Acrosome Genes." *Molecular Biology and Evolution* 27 (6): 1235–1246.

Dorus, Steve, Elaine C Wilkin, and Timothy L Karr. 2011. "Expansion and Functional Diversification of a Leucyl Aminopeptidase Family That Encodes the Major Protein Constituents of Drosophila Sperm." *BMC Genomics* 12: 177.

Fedorka, Kenneth M, Jodell E Linder, Wade Winterhalter, and Daniel Promislow. 2007. "Post-mating Disparity Between Potential and Realized Immune Response in *Drosophila melanogaster*." *Proceedings of the Royal Society B - Biological Sciences* 274 (1614): 1211–1217.

Findlay, Geoffrey D, and Willie J Swanson. 2010. "Proteomics Enhances Evolutionary and Functional Analysis of Reproductive Proteins." *BioEssays* 32 (1): 26–36.

Firman, Renée C, and Leigh W Simmons. 2011. "Experimental Evolution of Sperm Competitiveness in a Mammal." *BMC Evolutionary Biology* 11: 19.

Fuller, M T, J H Caulton, J A Hutchens, T C Kaufman, and E C Raff. 1988. "Mutations That Encode Partially Functional Beta 2 Tubulin Subunits Have Different Effects on Structurally Different Microtubule Arrays." *The Journal of Cell Biology* 107 (1): 141–152.

Gepner, J, and T S Hays. 1993. “A Fertility Region on the Y Chromosome of *Drosophila melanogaster* Encodes a Dynein Microtubule Motor.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (23): 11132–11136.

Goldstein, L S B, R W Hardy, and D L Lindsley. 1982. “Structural Genes on the Y Chromosome of *Drosophila melanogaster*.” *Proceedings of the National Academy of Sciences of the United States of America* 79 (23): 7405–7409.

Gomendio, Montserrat, and Eduardo R S Roldan. 2008. “Implications of Diversity in Sperm Size and Function for Sperm Competition and Fertility.” *The International Journal of Developmental Biology* 52 (5-6): 439–447.

Harris, Claire L, Masashi Mizuno, and B Paul Morgan. 2006. “Complement and Complement Regulators in the Male Reproductive System.” *Molecular Immunology* 43 (1-2): 57–67.

Hasan, A K M Mahbub, Yasuo Fukami, and Ken-ichi Sato. 2011. “Gamete Membrane Microdomains and Their Associated Molecules in Fertilization Signaling.” *Molecular Reproduction and Development* 78 (10-11): 814–830.

Hecht, N B. 1998. “Molecular Mechanisms of Male Germ Cell Differentiation.” *BioEssays* 20 (7): 555–561.

Hense, Winfried, John F Baines, and John Parsch. 2007. “X Chromosome Inactivation During Drosophila Spermatogenesis.” *PLoS Biology* 5 (10): 2288–2295.

Holt, William V, and Alireza Fazeli. 2010. “The Oviduct as a Complex Mediator of Mammalian Sperm Function and Selection.” *Molecular Reproduction and Development* 77 (11): 934–943.

Hosken, David J, and Paul I Ward. 2001. “Experimental Evidence for Testis Size Evolution via Sperm Competition.” *Ecology Letters* 4 (1): 10–13.

Hoyle, Henry D., and Elizabeth C. Raff. 1990. “Two Drosophila Beta Tubulin Isoforms Are Not Functionally Equivalent.” *The Journal of Cell Biology* 111 (3): 1009–1026.

Innocenti, P, and E H Morrow. 2009. “Immunogenic Males: a Genome-wide Analysis of Reproduction and the Cost of Mating in *Drosophila melanogaster* Females.” *Journal of Evolutionary Biology* 22 (5): 964–973.

Inoue, Naokazu, Masahito Ikawa, Tomoko Nakanishi, Misako Matsumoto, Midori Nomura, Tsukasa Seya, and Masaru Okabe. 2003. “Disruption of Mouse CD46 Causes an Accelerated Spontaneous Acrosome Reaction in Sperm.” *Molecular and Cellular Biology* 23 (7): 2614–2622.

Kaltschmidt, B, KH Glä tzer, F Michiels, D Leiss, and R Renkawitz-Pohl. 1991. "During Drosophila Spermatogenesis Beta 1, Beta 2 and Beta 3 Tubulin Isotypes Are Cell-type Specifically Expressed but Have the Potential to Coassemble into the Axoneme of Transgenic Flies." *European Journal of Cell Biology* 54 (1): 110–20.

Kapelnikov, Anat, Einat Zelinger, Yuval Gottlieb, Kahn Rhrissorrakrai, Kristin C Gunsalus, and Yael Heifetz. 2008. "Mating Induces an Immune Response and Developmental Switch in the Drosophila Oviduct." *Proceedings of the National Academy of Sciences of the United States of America* 105 (37): 13912–13917.

Karr, Timothy L. 2007. "Fruit Flies and the Sperm Proteome." *Human Molecular Genetics* 16 (2): 124–133.

Karr, Timothy L, and Scott Pitnick. 1996. "The Ins and Outs of Fertilization." *Nature* 379 (6564): 405–406.

Khil, Pavel P, Natalya A Smirnova, Peter J Romanienko, and R Daniel Camerini-Otero. 2004. "The Mouse X Chromosome Is Enriched for Sex-biased Genes Not Subject to Selection by Meiotic Sex Chromosome Inactivation." *Nature Genetics* 36 (6) (June): 642–646.

Kim, Hyeyeung, Rong Wu, Kathleen R Cho, Dafydd G Thomas, Gabrielle Gossner, J Rebecca Liu, Thomas J Giordano, Kerby A Shedden, David E Misek, and David M Lubman. 2008. "Comparative Proteomic Analysis of Low Stage and High Stage Endometrioid Ovarian Adenocarcinomas." *Proteomics. Clinical Applications* 2 (4): 571–584.

Kimble, Mary, Robert W Dettman, and Elizabeth C Raff. 1990. "The B3-Tubulin Gene of *Drosophila melanogaster* Is Essential for Viability and Fertility." *Genetics* 126 (4): 991–1005.

Klein, Theo, and Rainer Bischoff. 2011. "Active Metalloproteases of the A Disintegrin and Metalloprotease (ADAM) Family: Biological Function and Structure." *Journal of Proteome Research* 10 (1): 17–33.

Kleven, Oddmund, Terje Laskemoen, Frode Fossøy, Raleigh J Robertson, and Jan T Lifjeld. 2008. "Intraspecific Variation in Sperm Length Is Negatively Related to Sperm Competition in Passerine Birds." *Evolution* 62 (2): 494–499.

Krisfalusi, Michelle, Kiyoshi Miki, Patricia L Magyar, and Deborah A O'Brien. 2006. "Multiple Glycolytic Enzymes Are Tightly Bound to the Fibrous Sheath of Mouse Spermatozoa." *Biology of Reproduction* 75 (2): 270–278.

Lawniczak, Mara K N, Andrew I Barnes, Jon R Linklater, James M Boone, Stuart Wigby, and Tracey Chapman. 2007. "Mating and Immunity in Invertebrates." *Trends in Ecology & Evolution* 22 (1): 48–55.

Lawniczak, Mara K N, and David J Begun. 2004. “A Genome-wide Analysis of Courting and Mating Responses in *Drosophila melanogaster* Females.” *Genome* 47 (5): 900–910.

Leclerc, Vincent, and Jean-Marc Reichhart. 2004. “The Immune Response of *Drosophila melanogaster*.” *Immunological Reviews* 198: 59–71.

Li, Qing-Wei, Xiao-Yan Lu, Yong You, Huan Sun, Xin-Yu Liu, Jian-Zhong Ai, Rui-Zhi Tan, et al. 2012. “Comparative Proteomic Analysis Suggests That Mitochondria Are Involved in Autosomal Recessive Polycystic Kidney Disease.” *Proteomics* 12 (15-16): 2556–2570.

Mcgraw, Lisa A, Greg Gibson, Andrew G Clark, and Mariana F Wolfner. 2004. “Genes Regulated by Mating , Sperm , or Seminal Proteins in Mated Female *Drosophila melanogaster*.” *Current Biology* 14 (16): 1509–1514.

Meiklejohn, Colin D, Emily L Landeen, Jodi M Cook, Sarah B Kingan, and Daven C Presgraves. 2011. “Sex Chromosome-specific Regulation in the Drosophila Male Germline but Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation.” *PLoS Biology* 9 (8): e1001126.

Mhatre, Siddhita D, Brie E Paddock, Aleister J Saunders, and Daniel R Marenda. 2013. “Invertebrate Models of Alzheimer’s Disease.” *Journal of Alzheimer’s Disease* 33 (1): 3–16.

Mikhaylova, Lyudmila M, and Dmitry I Nurminsky. 2011. “Lack of Global Meiotic Sex Chromosome Inactivation, and Paucity of Tissue-specific Gene Expression on the Drosophila X Chromosome.” *BMC Biology* 9: 29.

Miles, Wayne O, Nicholas J Dyson, and James A Walker. 2011. “Modeling Tumor Invasion and Metastasis in Drosophila.” *Disease Models & Mechanisms* 4 (6): 753–761.

Miller, Gary T, and Scott Pitnick. 2002. “Sperm-female Coevolution in Drosophila.” *Science* 298 (5596): 1230–1233.

Mizuno, Masashi, Rossen M Donev, Claire L Harris, and B Paul Morgan. 2007. “CD55 in Rat Male Reproductive Tissue: Differential Expression in Testis and Expression of a Unique Truncated Isoform on Spermatozoa.” *Molecular Immunology* 44 (7): 1613–1622.

Mueller, Jacob L, Shantha K Mahadevaiah, Peter J Park, Peter E Warburton, David C Page, and James M A Turner. 2008. “The Mouse X Chromosome Is Enriched for Multicopy Testis Genes Showing Postmeiotic Expression.” *Nature Genetics* 40 (6): 794–799.

Nascimento, Jaclyn M, Linda Z Shi, James Tam, Charlie Chandsawangbhuwana, Barbara Durrant, Elliot L Botvinick, and Michael W Berns. 2008. “Comparison of Glycolysis and Oxidative Phosphorylation as Energy Sources for Mammalian Sperm Motility, Using the

Combination of Fluorescence Imaging, Laser Tweezers, and Real-time Automated Tracking and Trapping." *Journal of Cellular Physiology* 217 (3): 745–751.

Nixon, B, R J Aitken, and E A McLaughlin. 2007. "New Insights into the Molecular Mechanisms of Sperm-egg Interaction." *Cellular and Molecular Life Sciences* 64 (14): 1805–1823.

Noguchi, Tatsuhiko, Michiko Koizumi, and Shigeo Hayashi. 2011. "Sustained Elongation of Sperm Tail Promoted by Local Remodeling of Giant Mitochondria in Drosophila." *Current Biology* 21 (10): 805–814.

Oliva, Rafael, Sara de Mateo, and Josep Maria Estanyol. 2009. "Sperm Cell Proteomics." *Proteomics* 9 (4): 1004–1017.

Parisi, Michael, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. 2003. "Paucity of Genes on the Drosophila X Chromosome Showing Male-biased Expression." *Science* 299 (5607): 697–700.

Parker, GA. 1970. "Sperm Competition and Its Evolutionary Consequences in Insects." *Biol Rev Camb Philos Soc* 45: 525–67.

Peddinti, Divyaswetha, Bindu Nanduri, Abdullah Kaya, Jean M Feugang, Shane C Burgess, and Erdogan Memili. 2008. "Comprehensive Proteomic Analysis of Bovine Spermatozoa of Varying Fertility Rates and Identification of Biomarkers Associated with Fertility." *BMC Systems Biology* 2: 19.

Peng, Jing, Peder Zipperlen, and Eric Kubli. 2005. "Drosophila Sex-peptide Stimulates Female Innate Immune System After Mating via the Toll and Imd Pathways." *Current Biology* 15 (18): 1690–1694.

Pitnick, S, DJ Hosken, and TR Birkhead. 2009. "Sperm Morphological Diversity." In *Sperm Biology: An Evolutionary Perspective*, ed. Pitnick S Birkhead TR, Hosken DJ, 69–149. USA: Academic Press.

Pitnick, Scott, Therese Markow, and Greg S Spicer. 1999. "Evolution of Multiple Kinds of Female Sperm-Storage Organs in Drosophila." *Evolution* 53 (6): 1804–1822.

Prokupek, A M, S D Kachman, I Ladunga, and L G Harshman. 2009. "Transcriptional Profiling of the Sperm Storage Organs of *Drosophila melanogaster*." *Insect Molecular Biology* 18 (4): 465–475.

Qiu, Ning, Meihu Ma, Lei Zhao, Wen Liu, Yuqi Li, and Yoshinori Mine. 2012. "Comparative Proteomic Analysis of Egg White Proteins Under Various Storage Temperatures." *Journal of Agricultural and Food Chemistry* 60 (31): 7746–7753.

Razzell, William, Will Wood, and Paul Martin. 2011. “Swatting Flies: Modelling Wound Healing and Inflammation in Drosophila.” *Disease Models & Mechanisms* 4 (5): 569–574.

Rudrapatna, Vivek A, Ross L Cagan, and Tirtha K Das. 2012. “Drosophila Cancer Models.” *Developmental Dynamics* 241 (1): 107–118.

Stein, Kathryn K, Jowell C Go, William S Lane, Paul Primakoff, and Diana G Myles. 2006. “Proteomic Analysis of Sperm Regions That Mediate Sperm-egg Interactions.” *Proteomics* 6 (12): 3533–3543.

Sturgill, David, Yu Zhang, Michael Parisi, and Brian Oliver. 2007. “Demasculinization of X Chromosomes in the Drosophila Genus.” *Nature* 450 (7167): 238–242.

Swanson, W J, A G Clark, H M Waldrip-Dail, M F Wolfner, and C F Aquadro. 2001. “Evolutionary EST Analysis Identifies Rapidly Evolving Male Reproductive Proteins in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (13): 7375–7379.

Swanson, Willie J, Alex Wong, Mariana F Wolfner, and Charles F Aquadro. 2004. “Evolutionary Expressed Sequence Tag Analysis of Drosophila Female Reproductive Tracts Identifies Genes Subjected to Positive Selection.” *Genetics* 168 (3): 1457–1465.

Vibranovski, Maria D, Hedibert F Lopes, Timothy L Karr, and Manyuan Long. 2009. “Stage-specific Expression Profiling of Drosophila Spermatogenesis Suggests That Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-expressed Genes.” *PLoS Genetics* 5 (11): e1000731.

Wasbrough, Elizabeth R, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L Karr. 2010. “The *Drosophila melanogaster* Sperm proteome-II (DmSP-II).” *Journal of Proteomics* 73 (11): 2171–2185.

Xue, Kun, Jing Yang, Biao Liu, and Dayuan Xue. 2012. “The Integrated Risk Assessment of Transgenic Rice Oryza Sativa: A Comparative Proteomics Approach.” *Food Chemistry* 135 (1): 314–318.

Chapter 3

Retrogenes contribute to disparate metabolic processes in mammalian and Drosophila sperm

3.1 Introduction

Spermatozoa have the conserved function of delivering the paternal haploid genome to the oocyte in sexually reproducing species. Despite this critical function, sperm are amongst the fastest evolving cell types and display huge morphological and physiological diversity across species (Pitnick, Hosken, and Birkhead 2009; Gage 2012). This rapid diversification of sperm is believed to be the result of sexual selection, mediated by sperm competition (Pitnick, Hosken, and Birkhead 2009; Gage 2012). Sperm competition occurs when sperm from rival males compete to fertilise a single oocyte (Parker 1970). Sperm competition can affect both sperm quantity and quality (Hosken and Ward 2001; Morrow and Gage 2001; Gomendio and Roldan 2008; Kleven et al. 2008; Firman and Simmons 2011). Sperm quantity can be experimentally selected by varying the intensity of sperm competition over several generations. For example, males of *Mus domesticus* (house mouse) produce ejaculates with increased sperm numbers and increased sperm motility when kept under polygamous compared to monogamous conditions (Firman and Simmons 2011). In addition to changes in sperm numbers, sperm morphology can be affected by altering the intensity of sperm competition. For instance, lines of *Caenorhabditis elegans* selected under high sperm competition produced sperm with 20% larger volume than control lines (LaMunyon and Ward 2002). As sperm from different species display large differences in both number and morphology, and as males from different species have different reproductive strategies, it is likely that there are a wide variety of mechanisms for improving sperm competitive ability, and it is these mechanisms that are under intense sexual selection.

The only system-level analysis of sperm proteome evolution to date involved the comparison of the *Drosophila melanogaster* and *M. musculus* (mouse) sperm proteomes, which was presented in the proceeding chapter (Rettie and Dorus 2012). This analysis revealed high levels of functional similarities, including similar proportions of genes involved in metabolic process, cell component organisation and the cell cycle. However, the spermatozoa from these species are morphologically and physiologically distinct, and it is possible that signatures of these differences may be detectable at the genomic level; for example, processes enriched with newly created genes or genes evolving under positive selection.

Mammalian and Drosophila sperm competitive ability is determined by different biological traits, which are under differing post-copulatory selection regimes. Mammalian spermatozoa tend to be small highly motile cells with three distinct morphological sections: head, midpiece and flagellum. In mammals, midpiece volume is positively correlated with both testis size and the degree of polyandry between species (Anderson, Nyholt, and Dixson 2005). In addition, the midpiece has been determined to be the only morphological indicator of sperm swimming speed in house mice (Firman and Simmons 2011). In many species including Iberian red deer (*Cervus elaphus hispanicus*) (Malo et al. 2005), and Atlantic salmon (*Salmo salar*) (Gage et al. 2004) sperm velocity is positively correlated with sperm fitness and share of paternity. However, despite the dense packaging of mitochondria into the midpiece, mitochondrial derived energy for sperm motility has been called into question, primarily due to the difficulties in diffusing ATP from the midpiece to the distal end of the flagellum (Nascimento et al. 2008). Furthermore, there is evidence that ATP required for sperm motility is predominantly generated in the flagellum by glycolytic enzymes distributed along the fibrous sheath of the axoneme (Krisfalusi et al. 2006). Due to the importance of sperm metabolic pathways for sperm motility it is likely that processes associated with this function are under intense selection.

Unlike mammals, sperm velocity does not appear to be a primary predictor of the outcome of sperm competition in Drosophila. Within the Drosophila subgenus sperm gigantism is common, and sperm length varies by two orders of magnitude, ranging from ~0.36mm in *D. pseudoobscura* to ~60.00mm in *D. bifurca*, with *D. melanogaster* spermatozoa ~1.91mm (Karr and Pitnick 1996). Furthermore, *D. melanogaster* males that have been selected for longer sperm performed better than males selected for short sperm in sperm competition experiments (Miller and Pitnick 2002). In addition, there is strong experimental evidence demonstrating that this trait is co-evolving with the female reproductive tract (Miller and Pitnick 2002). In the Drosophila genus the evolution of sperm gigantism, and the competitive advantage long sperm provide, may be intimately associated with the evolution of the nebenkern (Noguchi, Koizumi, and Hayashi 2012). During late meiosis Drosophila mitochondria fuse, and these giant mitochondrial derivatives are packaged into the structure termed the nebenkern (Noguchi, Koizumi, and Hayashi 2011). Finally, during post-meiosis the nebenkern restructures in parallel with the developing axoneme. Disruption of the nebenkern-microtubule linkage results in defects in uniaxial spermatid tail development and male infertility (Noguchi, Koizumi, and Hayashi 2011). Therefore, while sexual selection in mammalian sperm is likely intense on processes related to sperm motility, in Drosophila it is probable that sexual selection is primarily focused on processes associated with sperm tail development.

In addition to the overall conservative evolution of sperm genes (Dorus et al. 2006; Dorus et al. 2010; Wasbrough et al. 2010) and the functional conservation between Drosophila and mouse proteomes (Rettie and Dorus 2012), gene duplication has been implicated as an important mechanism in sperm evolution (Dorus et al. 2008). Gene duplication is an essential mechanism of evolution as it results in the creation of new genes via DNA- or RNA-based mechanisms (Lynch and Conery 2003; Kaessmann 2010). In both Drosophila and mammals, gene duplication has contributed to the creation of genes with male-biased expression or function (male-biased genes), including genes implicated in spermatogenesis and sperm evolution. The analysis of the *D. melanogaster* sperm proteome (DmSP) revealed an abundance of both retrogenes and tandem gene duplicates encoding novel sperm proteins (Dorus et al. 2008). These included the retrogene *CG13340* (*S-LAP7*), a member of the sperm leucyl aminopeptidase (S-LAP) gene family. All seven members of the S-LAP family have been identified in the DmSP and are the most abundant proteins, by mass, in the sperm proteome (Wasbrough et al. 2010). The expansion of this gene family is proposed to have involved one retrotransposition event and two independent tandem DNA-based gene duplication events. Although each S-LAP contains a catalytic domain, the enzyme activity of the ancestral gene has been lost, and the current role of these genes in sperm is unclear (Dorus, Wilkin, and Karr 2011). Sequential gene duplication has also led to the formation of two X-linked gene clusters in Drosophila: *Sdic* and *tektin*, both of which have functions important to sperm competitive ability (Greenspan and Clark 2011; Yeh et al. 2012). In mammals, frame shift mutations in the retrotransposed mouse gene *Utp14b* resulted in the arrest of spermatogenesis (Bradley et al. 2004), while a single nucleotide polymorphism in the human retrogene *PGAM4* has been linked with human infertility (Okuda et al. 2012). Finally, in both mammals and Drosophila duplication of genes associated with metabolic pathways has occurred. In *D. melanogaster,* nuclear encoded mitochondrial genes have been duplicated (Gallach, Chandrasekaran, and Betrán 2010). In mammals, several glycolytic enzymes have produced retrotransposed gene copies (Vemuganti, De Villena, and O'Brien 2010). In both taxa, these duplicates have been proposed to function in sperm (Vemuganti et al. 2007; Gallach, Chandrasekaran, and Betrán 2010; Vemuganti, de Villena, and O'Brien 2010). These individual cases show that retrotransposition has been an important mechanism in the creation of novel sperm genes. However, while it is well documented that retrogenes often acquire testis expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004), the extent to which retrotransposition contributed to the evolution of the sperm proteome has not been characterised. Little is known about the function of retrogenes in spermatozoa, or if they are associated with adaptive changes in sperm biology. In this study, we conducted a survey of all retrogenes encoding sperm components in Drosophila and mammals, in order to assess the contribution of

retrotransposition to processes related to post-copulatory selection of sperm in taxa where pronounced differences in post-copulatory selection on sperm exist.

3.2 Materials and methods

3.2.1 Characterisation of retrogenes

D. melanogaster (hereafter referred to as Drosophila) retrogenes were curated from previously characterised retrogene datasets (Bai et al. 2007; Zhou et al. 2008). Similar methods to those used in Drosophila were employed to conservatively characterise retrogenes in three mammalian taxa (Betrán, Thornton, and Long 2002). RefSeq genes and their exon counts were downloaded from UCSC (http://genome.ucsc.edu/) based on the genome builds NCBI37 (*M. musculus*), Baylor (*Rattus norvegicus*) and GRCh3 (*Homo sapiens*) and corresponding nucleotide (protein) coding sequences were obtained from NCBI (http://www.ncbi.nlm.nih.gov/). tBLASTx was used to identify paralogous genes within each respective genome. Putative retrogenes were identified as single exon genes whose closest paralog in the genome was a multi-exon gene with which it shared ≥ 70% sequence identity in an alignment including ≥ 70% of the gene sequence. Genes where a single parental gene could not be definitively identified were not used.

3.2.2 Timing of retrotransposition events

A comparative genomic approach was used to infer the timing of retrotransposition events using annotated orthology databases from MGI (http://www.informatics.jax.org/) and FlyBase (www.flybase.org). In Drosophila, this included a range of species across the Drosophilidae family (Figure 3.1). The presence/absence of annotated orthologs for each retrogene was used to parsimoniously define the lineage where each retrogene was first observed. The dog (*Canis lupus familiaris*) and the mosquito (*Anopheles gambiae*) genomes were used as outgroups to determine which retrogenes predated the divergence of the Euarchontoglires and Drosophilidae groups, respectively.

3.2.3 Identification of sperm retrogenes

Drosophila sperm retrogenes were defined as retrogenes that encode a sperm protein previously characterised by mass spectroscopy (MS) in the *D. melanogaster* sperm proteome (Dorus et al. 2006; Wasbrough et al. 2010). Similarly, mammalian sperm retrogenes were any retrogenes that encoded a previously characterised sperm protein in either the mouse, rat or human sperm proteome (*M. musculus* (Cao, Gerton, and Moss 2006; Stein et al. 2006; Baker, Hetherington, Reeves, and Aitken 2008a; Dorus et al. 2010), *R. norvegicus* (Baker et al. 2008b) and/or *H. sapiens* (Baker et al. 2007)).

3.2.4 Analysis of sperm retrogene function

Drosophila and mammalian sperm retrogene function was investigated based on gene ontology (GO) functional category annotation. In mammals, mouse retrogenes, or the closest mouse

paralog for retrogenes found exclusively in the other mammalian species studied, were used to query the GO database. GO biological process enrichment was determined using a hypergeometric distribution with Bonferroni correction for multiple testing, as implemented by GOEAST (http://omnicslab.genetics.ac.en/GOEAST). Sperm proteome genes with functions in metabolic processes (GO:0008152), Carbohydrate metabolic process (GO:0005975) and Oxidation-reduction processes (GO:0008152) were determined based on their presence in the gene annotation files obtained from AmiGO (http://www.geneontology.org). To determine whether enrichment/deficiency of sperm retrogenes functioning in metabolic process, or within specific metabolic pathways, was due to similar enrichments/deficiencies in the composition of the sperm proteomes, the expected number of sperm retrogenes based on the proportion of genes with these functions in the sperm proteome was calculated. The percentage excess (or deficiency) of sperm retrogenes within each category compared to the percentage of the sperm proteome within each category was also calculated. Significance was determined using a two-tailed χ^2 test, without Yates correction. Weighted binomial probabilities, based on the proportion of the sperm proteome within these GO categories, were used to determine the likelihood of the sperm proteome generating the observed number of sperm retrogenes with functions in each GO category. An identical analysis was conducted to determine whether the enrichment of sperm retrogenes in these three categories was due to the composition of the respective genome.

3.2.5 Analysis of interchromosomal retrogene movements

The direction of interchromosomal retrotransposition in mammals was inferred as the most parsimonious explanation of the parent and retrogene annotated locations in the three mammalian genomes. In Drosophila, this was based on parent-retrogene location in *D. melanogaster*. The expected number of X-to-autosome and autosome-to-autosome retrotransposition events was determined based on the methodology of Betran *et al*. (2002). This formula accounts for differences in chromosome size and chromosome gene number and the population size of the X chromosome compared to the autosomes. A χ^2 test was used to detect significant disparities between the observed and expected movements of sperm retrogenes. An identical analysis was conducted to statistically analyze the movement of retrogenes with specific functions in relation to the physical distribution of similar functioning genes across the entire genome or those genes that encode the Drosophila or mouse sperm proteins. This included genes functioning in metabolic processes (GO:0008152), carbohydrate metabolic process (GO:0005975) and oxidation-reduction processes (GO:0055114) based on gene annotation files in AmiGO (http://www.geneontology.org).

3.2.6 Analysis of sperm retrogene expression patterns

Microarray expression data for testis and non-reproductive organs was obtained from Gene Atlas (probe set GNF1M; http://biogps.org/) and FlyAtlas (http://www.flyatlas.org/) (Chintapalli, Wang, and Dow 2007) for mammals and Drosophila, respectively. Average somatic expression in mammals was calculated using expression levels in the cerebellum, heart, hypothalamus, kidney, liver, lung, pancreas and skeletal muscle. Gene expression in the mouse for the three primary stages of spermatogenesis (mitosis, meiosis and post-meiosis) was based on microarray data from four cell types associated with these states, Type A spermatogonia, Type B spermatogonia, Pachytene spermatocytes and round spermatids (http://mrg.genetics.washington.edu), In Drosophila, stage-specific gene expression during spermatogenesis was based on microarray analysis of samples enriched for cells in the mitotic, meiotic and post-meiotic stages of development (Vibranovski et al. 2009). Parent-retrogene gene pairs were excluded from these analyses if microarray probe sequences did not differentiate between the parent and the retrogene nucleotide sequences. Difference in expression between each stage of spermatogenesis was assessed using non-parametric Kolmogorov-Smirnov tests. Average expression in the testis, non-reproductive tissues and throughout spermatogenesis was statistically assessed using Kolmogorov-Smirnov tests between sperm retrogenes and their parents, and between non-sperm retrogenes and sperm retrogenes.

3.3 Results

3.3.1 Retrotransposition has contributed to the creation of novel sperm proteins

Characterisation of retrogenes in the human, mouse and rat genomes resulted in the identification of 135 mammalian retrotransposition events, of which 25 (18.5%) encode sperm proteins based on MS identification (Appendix V). A total of 101 retrogenes were documented in Drosophila using the same methodology. Twenty of these retrogenes (19.8%) encode components of the sperm proteome (Appendix VI), a proportion that is statistically indistinguishable ($\chi^2 = 0.007$, $p = 0.9355$) from the proportion of sperm retrogenes observed in mammals. It is also noteworthy that the majority of the Drosophila sperm retrogenes originated before the diversification of the Drosophila genus (Figure 3.1), while mammalian sperm retrogenes have originated in a more consistent manner across the analysed lineages within the mammalian phylogeny (Figure 3.2).

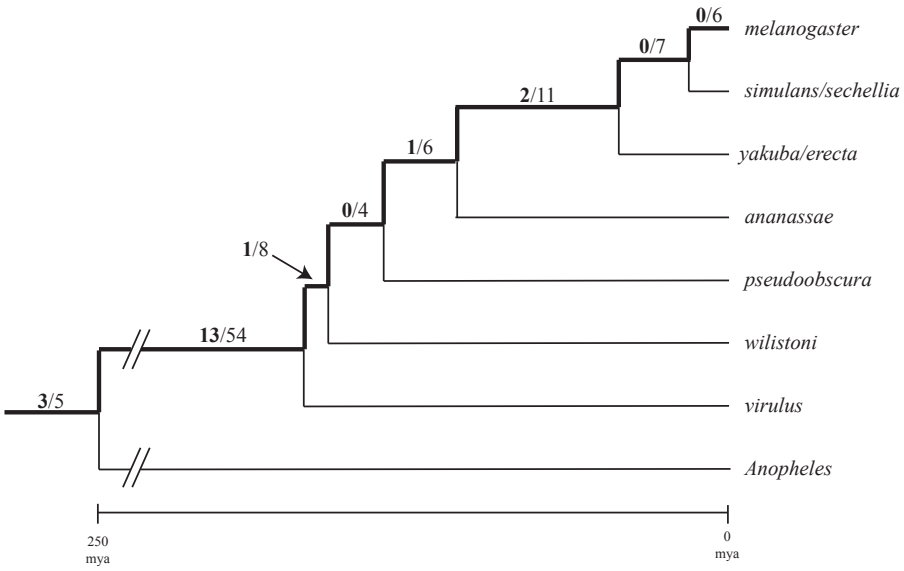

Figure 3.1 Retrotransposition events in the Drosophilidae family.

Numbers above lineages denote the number of sperm retrogenes (bold) out of the total number of retrotransposition events that occurred on the lineage.

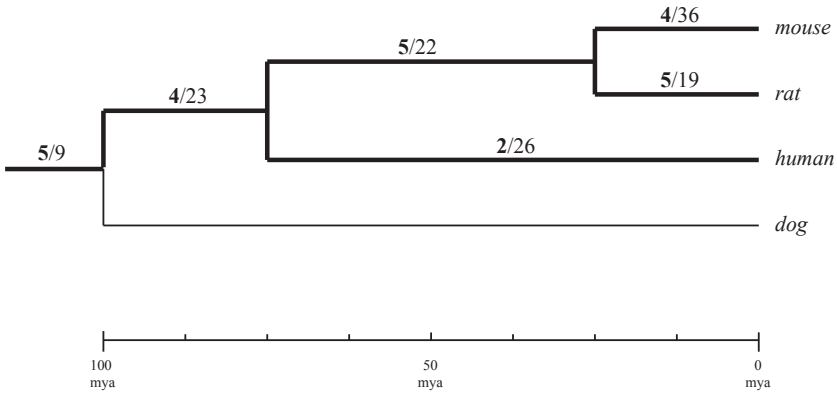

Figure 3.2 Retrotransposition events in the Euarchontoglires clade.
Numbers above lineages denote the number of sperm retrogenes (bold) out of the total number of retrotransposition events that occurred on the lineage.

3.3.2 Sperm retrogenes are enriched in metabolic pathways

Gene Ontology analysis of biological process revealed an enrichment of sperm retrogenes functioning in metabolic processes in both mammals ($p < 0.0001$) and Drosophila ($p = 0.0017$) (Appendices IV and V). The proportion of sperm retrogenes with a known function in metabolism (>70%) was statistically indistinguishable between mammals and Drosophila ($\chi^2 = 0.051$; $p = 0.8211$). Further, we also observed several examples of sperm retrogenes that likely function in related processes within both sperm proteomes. These include the rodent-specific retrogene *Pbp2*, which is the progeny of *Pebp1* that is involved in the removal of cell surface proteins that inhibit capacitation (Gibbons, Adeoya-Osiguwa, and Fraser 2005) and the putative Drosophila phosphatidylethanolamine binding protein (PEBP) the sperm retrogene *CG6180* (Rautureau et al. 2009). Additionally, the creation of testis proteasome retrogenes is abundant in both Drosophila and mammals. In Drosophila, the retrogene *Prosα6T* (Belote and Zhong 2009) was identified in the DmSP. While in mammals, the sperm retrogene *RGD1560350* is a derivative of the proteasome component *Psma6*. Mammalian sperm retrogenes of regulators of proteasome precursors, including *Rhoa* and *Rhoc* (Zhang et al. 2010), were also found. In addition to functions in metabolism, several Drosophila sperm retrogenes with known or potential functions in sperm development, including *gskt*, which functions in male gamete generation (Kalamegham et al. 2007), *Hsp60B*, which is involved in spermatid development (Castrillon et al. 1993), and *Prosα6T*, which is involved in spermatid nucleus elongation and

sperm individualisation (Zhong and Belote 2007), were found. All three retrogenes have male sterile mutants (www.flybase.org). Finally, we also identified genes that may have a structural role in Drosophila sperm, including *Act87E*, which functions in cytoskeleton organisation, and *Cdlc2*, which has a role in microtubule-based movements.

3.3.3 Sperm retrogenes are enriched in disparate metabolic pathways in mammals and Drosophila

Gene ontology enrichment analysis revealed that while sperm retrogenes from both taxa were enriched with metabolic process functions, mammalian and Drosophila sperm retrogenes were enriched with functions from different metabolic pathways (Figure 3.3). Mammalian metabolic sperm retrogenes are primarily enriched for functions in carbohydrate metabolism. In addition to previously identified mammalian glycolytic retrogenes (*Pgk2, Aldoal1, Aldoart1*) (Vemuganti, De Villena, and O'Brien 2010), an enrichment of mammalian sperm retrogenes functioning in phosphorylation was revealed; including *Gk2, Gykl1, Prps1l1* and two lineage-specific kinases: human *PRKACG* and rat *Prkar1a* (Appendix V). A further two mammalian sperm retrogenes *G6pda2* and *PRPS1L1*, were identified within the pentose phosphate pathways. In contrast, Drosophila metabolic sperm retrogenes are enriched for metabolic energetic pathways associated with mitochondrial energy production, including the tricarboxylic acid cycle, electron transport and oxidative phosphorylation (OXPHOS). This is consistent with previously documented retrogenes with functions in mitochondrial energetic pathways (Gallach, Chandrasekaran, and Betrán 2010).

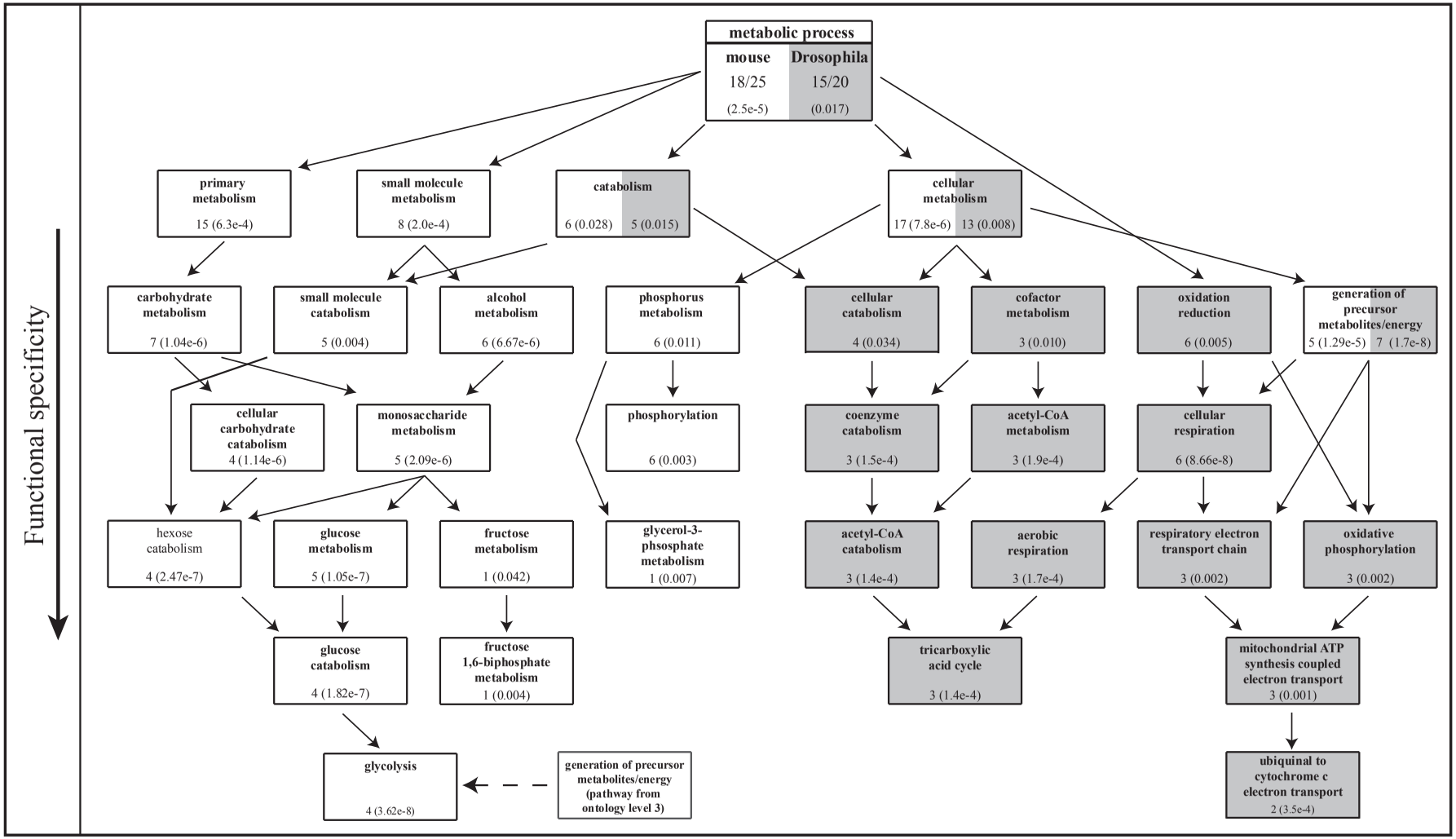

Figure 3.3 Metabolic functions of mammalian and *Drosophila* sperm retrogenes.

Gene Ontology biological processes hierarchy diagram where all categories are required to be significant ($p<0.05$) and form an uninterrupted pathway from ontology level 1. Significant enrichment of Drosophila (grey shading) and mouse (white) retrogenes per category is indicated, including the number of retrogenes and the level of significance.

3.3.4 Genome composition does not explain enrichment of metabolic sperm retrogenes

To determine whether the enrichment of metabolic processes within mammalian and Drosophila sperm retrogenes was due to the GO composition of their respective genomes, we calculated the proportion of genes with functions in metabolic processes in the genome (Table 3.1). We observed that a significantly higher proportion of sperm retrogenes function in metabolism compared to the whole genome in both mammals (χ^2 = 25. 587; $p < 0.0001$) and Drosophila (χ^2 = 9.232; $p = 0.0024$). Weighted binomial distributions revealed that there are significantly more metabolic sperm retrogenes than expected based on the whole genome composition in Drosophila ($n = 20$, $k = 15$, $q = 0.6071$; probability exactly 15 out of 20 = 0.0010) or mammals ($n = 25$, $k = 18$, $q = 0.7426$; probability exactly 18 out of 25 < 0.0001). In addition, we observed a significantly higher proportion of mammalian (χ^2 = 26.10; $p < 0.0001$), but not Drosophila (χ^2 = 0.271; $p = 0.6029$), metabolic sperm retrogenes function in carbohydrate metabolism compared to metabolic genes in the genome. In mammals, the number of metabolic sperm retrogenes with functions in carbohydrate metabolism is significantly higher than expected based on genome composition ($n = 18$, $k = 7$, $q = 0.0642$; probability exactly 7 out of 18 < 0.0001), in contrast to Drosophila, which lacks significant over-abundance of carbohydrate metabolism retrogenes ($n = 15$, $k = 2$, $q = 0.9336$; probability exactly 2 out of 15 = 0.1895). Conversely, we observed a significantly higher proportion of Drosophila (χ^2 = 11.694; $p = 0.0006$), but not mammalian (χ^2 = 1.315; $p = 0.2516$), metabolic sperm retrogenes function in oxidative-reduction process compared to metabolic genes in general. The number metabolic sperm retrogenes that function in oxidation-reduction processes is significantly greater than expected in Drosophila ($n = 15$, $k = 6$, $q = 0.8998$; probability exactly 6 out of 15 = 0.0020), while in mammals they are not significantly over-abundant ($n = 18$, $k = 3$, $q = 0.9302$; probability exactly 3 out of 18 = 0.0937).

Table 3.l Enrichment of metabolic sperm retrogenes relative to the genome

	Sperm retrogenes	Genome	Retrogene excess (%)	p value
Metabolic processes (GO:0008152)				
Mammals	18/25	8910/34621	+46.26	<0.0001
Drosophila	15/20	5709/14532	+ 35.71	0.0024
Carbohydrate metabolism (GO:0005975)				
Mammals	7/18	572/8910	+32.47	<0.0001
Drosophila	2/15	379/5709	+6.69	0.6029
Mitochondrial energetics (Oxidation-reduction process: GO:0055114)				
Mammals	3/18	622/8910	+9.69	0.2516
Drosophila	6/15	572/5709	+29.98	0.0006

p values based on χ^2 comparison of the proportion of genes in each GO category

3.3.5 Sperm proteome composition does not explain enrichment of metabolic sperm retrogenes

Similarly, to determine whether the enrichment of metabolic processes within mammalian and Drosophila sperm retrogenes was a representation of the underling GO composition of their proteomes, we calculated the proportion of each proteome with functions in metabolic processes (Table 3.2). We observed that a significantly higher proportion of sperm retrogenes have functions in metabolic processes than in the sperm proteome as a whole, in Drosophila (χ^2 = 7.062; p = 0.0079) but not mammals (χ^2 = 2.725; p = 0.0988). Weighted binomial distributions revealed that there are significantly more metabolic sperm retrogenes than expected based on the whole sperm proteome composition in Drosophila (n = 20, k = 15, q = 0.5726; probability exactly 15 out of 20 = 0.0028) and mammals (n = 25, k = 18, q = 0.4671; probability exactly 18 out of 25 = 0.0280). Similar to comparisons with the whole genome, we observed a significantly higher proportion of mammalian (χ^2 = 8.780; p = 0.0030), but not Drosophila (χ^2 = 0.007; p = 0.9328), metabolic sperm retrogenes function in carbohydrate metabolism compared to metabolic genes in the sperm proteome. In mammals, the number of metabolic sperm retrogenes with functions in carbohydrate metabolism is significantly higher than expected based on sperm proteome composition (n = 18, k = 7, q = 0.8788; probability exactly 7 out of 18 = 0.0030), in contrast to Drosophila, which lacks significant over-abundance of carbohydrate metabolism retrogenes (n = 15, k = 2, q = 0.9075; probability exactly 2 out of 15 = 0.2544). Conversely, we did not observe a significantly higher proportion of Drosophila (χ^2 = 2.199, p =

0.1381) or mammalian (χ^2 =0.030; p = 0.8625) metabolic sperm retrogenes function in oxidative-reduction process compared to metabolic sperm proteome genes. The number metabolic sperm retrogenes that function in oxidation-reduction processes is significantly greater than expected in Drosophila (n = 15, k = 6, q = 0.7935; probability exactly 6 out of 15 = 0.0484), while in mammals they are not significantly over-abundant (n = 18, k = 3, q = 0.8173; probability exactly 3 out of 18 = 0.2413).

Table 3.2 Enrichment of metabolic sperm retrogenes relative to the sperm proteome

	Sperm retrogenes	Sperm proteome	Retrogene excess (%)	p value
Metabolic processes (GO:0008152)				
Mammals	18/25	520/979	+18.7	0.0988
Drosophila	15/20	465/1088	+32.3	0.0079
Carbohydrate metabolism (GO:0005975)				
Mammals	7/18	63/520	+26.8	0.0030
Drosophila	2/15	43/465	+4.1	0.9328
Mitochondrial energetics (Oxidation-reduction process: GO:0055114)				
Mammals	3/18	95/520	-1.6	0.8625
Drosophila	6/15	96/465	+19.4	0.1381

p values based on χ^2 comparison of the proportion of genes in each GO category

3.3.6 Excess movement of mammalian, but not Drosophila, sperm retrogenes from the X chromosome

We observed no X-linked sperm retrogenes in either mammals or Drosophila. As an excess of retrogene movement off the X chromosome has been observed in both mammals (Emerson et al. 2004) and Drosophila (Betrán, Thornton, and Long 2002) we sought to confirm this pattern for sperm retrogenes. We observed a significant excess of sperm retrogenes resulting from X to autosome movements in mammals (p = 0.01808) but not Drosophila (p = 0.1247) (Table 3.3).

Table 3.3 Expected X chromosome to autosome movement based on whole genome

Taxa	# Expected	# Observed	p value
Mammals	1	10	0.01808
Drosophila	4	5	0.1247

To determine whether the observed differences in the number of sperm retrogenes relocating from the X chromosome to the autosomes is due the underlying genomic distribution of sperm proteome genes we repeated the analysis using only sperm proteome genes. Similar to the analysis with all genes in the genome, we observed a significant excess of mammalian ($p = 0.0046$), but not Drosophila ($p = 0.2471$), sperm retrogenes resulting from X to autosome movements based on the genomic distribution of MmSP and DmSP genes, respectively (Table 3.4).

Table 3.4 Expected X chromosome to autosome movement based on sperm proteome distribution

Taxa	# Expected	# Observed	p value
Mammals	1	10	0.0046
Drosophila	5	5	0.2471

3.3.7 Excess movement of mammalian, but not Drosophila, metabolic sperm genes from the X chromosome

To determine whether there was a similar excess of X to autosome relocations for metabolic sperm retrogenes we determined the expected number of X to autosome movements based on the distribution of metabolic genes in the genome. We observed a significant excess of metabolic sperm retrogenes relocating from the X chromosome to the autosomes in mammalian ($p = 0.0043$), but not in Drosophila ($p = 0.5680$) (Table 3.5). However, when we considered only those genes with the GO category carbohydrate metabolism we observed a non-significant excess of mammalian ($p = 0.0699$) carbohydrate metabolism retrogenes relocating from the X chromosome, and no Drosophila carbohydrate sperm retrogenes retroposed from the X chromosome to the autosomes. Similarly, when we considered only those genes with oxidative-reduction process functions, we observed no excess movement of mammalian or Drosophila mitochondrial sperm retrogenes retroposed from the X chromosome.

Table 3.5 Expected X chromosome to autosome movement, based on distribution of metabolic genes in genome

Taxa	# Expected	# Observed	p value
Metabolic processes (GO:0008152)			
Mammals	1	9	0.0043
Drosophila	3	3	0.5680
Carbohydrate metabolism (GO:0005975)			
Mammals	0	4	0.0699
Drosophila	0	0	1.0000
Mitochondrial energetics (Oxidation-reduction process: GO:0055114)			
Mammals	0	1	1.0000
Drosophila	2	1	1.0000

We also considered whether the genomic distribution of metabolic sperm proteome genes might explain the different patterns of X to autosome movement of Drosophila and mammalian sperm retrogenes. When we considered only metabolic sperm proteome genes we observed a significant excess of mammalian ($p = 0.0043$), but not Drosophila ($p = 0.3633$), metabolic sperm retrogenes moving from the X chromosome to the autosomes (Table 3.6). However, when we considered only sperm proteome genes with functions in carbohydrate metabolism, or mitochondrial process, we observed no significant excess of sperm retrogenes, within these GO categories, retrotransposed from the X chromosome.

3.3.8 The majority of mammalian sperm retrogenes have parental genes that also encode sperm components

We observed that a significantly higher proportion of parental genes also encode sperm proteins in mammals (76%; 19 out of 25) than in Drosophila (15%; 3 out of 20) ($p < 0.0001$). Of the 3 Drosophila sperm retrogenes with parents that also encode sperm proteins, all were the result of autosome-to-autosome movements, and all had an annotated function in metabolism. For the 19 mammalian sperm retrogenes with parents encoding sperm proteins, 7 out of 19 were the result of X to autosome retrotranspositions, while the remaining 11 out of 19 were the result of autosome-to-autosome movements. The majority of these mammalian sperm retrogenes had functions in metabolism, including 6 out of the 7 X-to-autosome and 7 out of 11 autosome-to autosome sperm retrogenes.

Table 3.6 Expected X chromosome to autosome movement, based on the genomic distribution of metabolic sperm proteome genes

Taxa	# Expected	# Observed	p value
Metabolic processes (GO:0008152)			
Mammals	1	9	0.0043
Drosophila	4	3	0.3633
Carbohydrate metabolism (GO:0005975)			
Mammals	0	4	0.0699
Drosophila	0	0	1.0000
Mitochondrial energetics (Oxidation-reduction process: GO:0055114)			
Mammals	0	1	1.0000
Drosophila	2	1	0.5455

3.3.9 Analysis of sperm retrogene tissue expression

Consistent with previous studies (Betrán, Thornton, and Long 2002; Emerson et al. 2004), sperm retrogenes are expressed at significantly higher levels in the whole testis relative to somatic tissues in both taxa (Figure 3.4). Similarly, sperm retrogenes are expressed at significantly lower levels in somatic tissues compared to their parental genes. However, Drosophila sperm retrogenes are expressed at significantly higher levels in the testis compared to their parental genes, while mammalian sperm retrogenes are expressed in the testis at comparable levels to their parental genes.

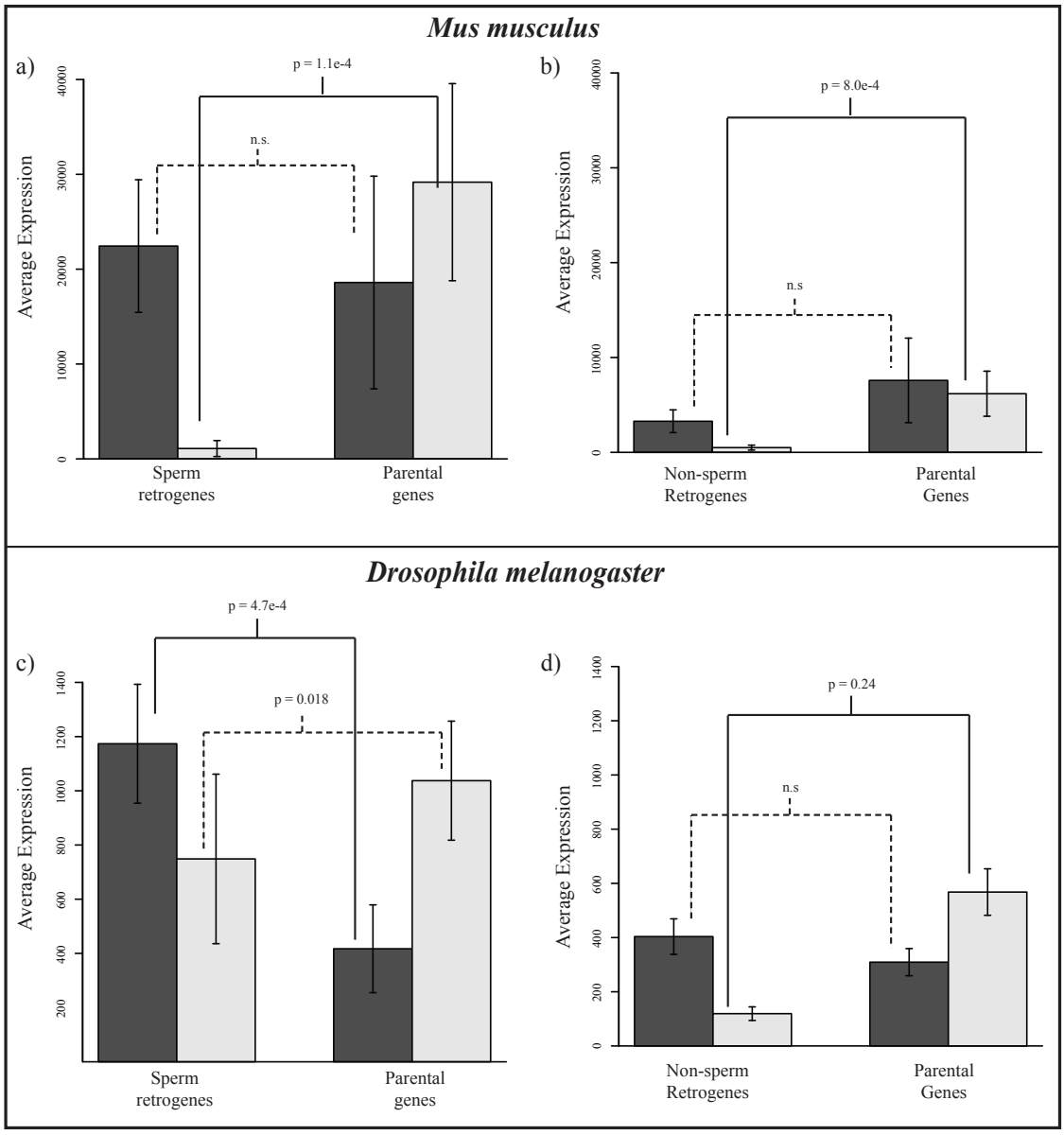

Figure 3.4 Testis and somatic tissue expression of retrogenes and their parent genes.

Average expression in the testis (dark gray) and somatic tissue average (light gray) for (a) mouse sperm retrogenes, (b) mouse non-sperm retrogenes, (c) Drosophila sperm retrogenes and (d) Drosophila non-sperm retrogenes and their respective parent genes. Expression data was obtained from GeneAtlas (http://biogps.org/) and FlyAtlas (Chintapalli, Wang, and Dow 2007) for mouse and Drosophila, respectively. Standard errors and significance of expressional differences, based on non-parametric Kolmogorov-Smirnov tests, between retrogenes and parental genes are indicated

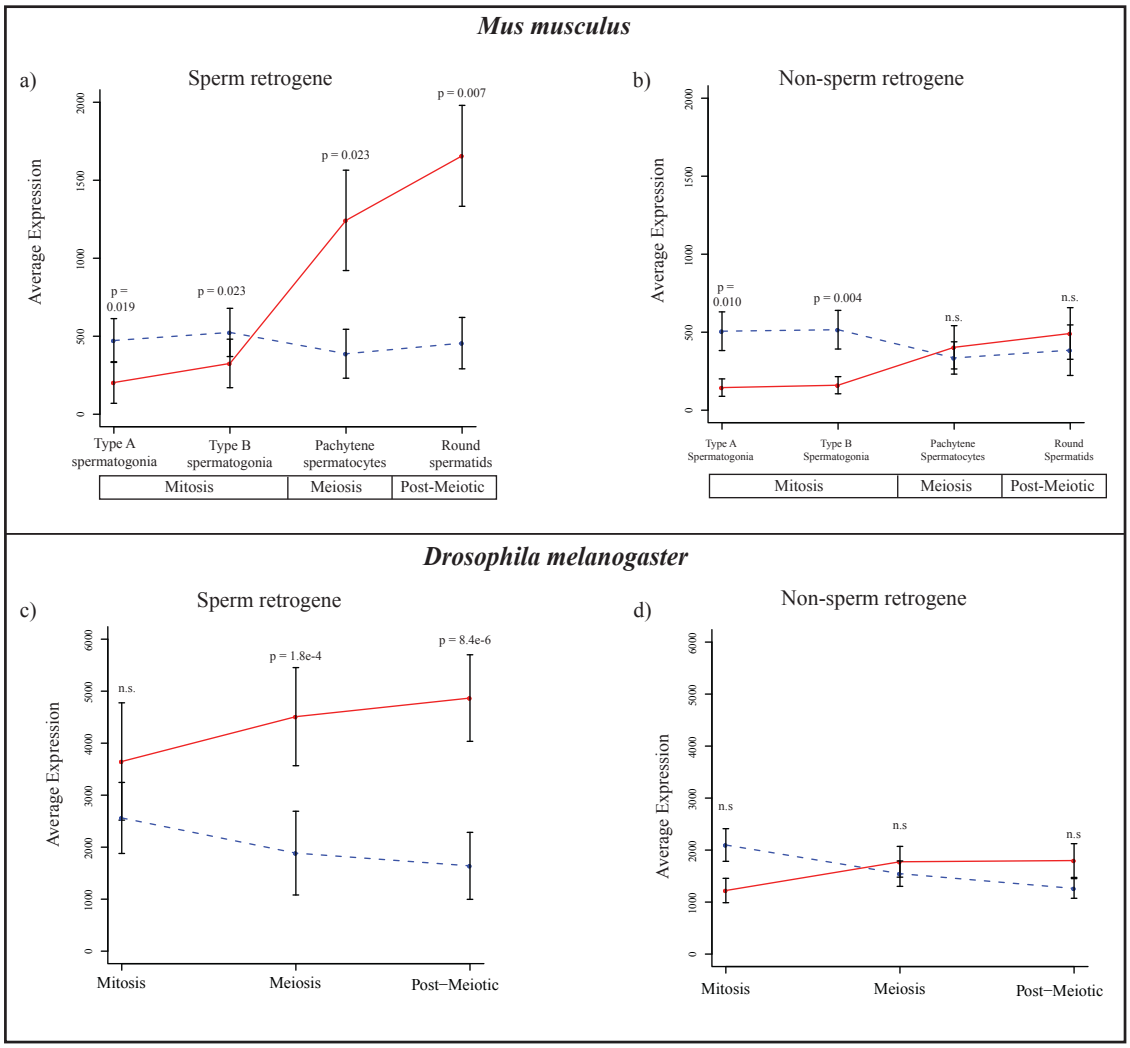

Figure 3.5 Retrogene and parent gene expression during spermatogenesis.

Average retrogene (solid, red) and parental (dotted, blue) expression during the spermatogenic stages is displayed for (a) mouse sperm retrogenes, (b) mouse non-sperm retrogenes, (c) Drosophila sperm retrogenes and (d) Drosophila non-sperm retrogenes and their parent genes. Expression data was obtained from Mammalian Reproductive Genetics (http://mrg.genetics.washington.edu) and Vibranovski et al (2009) for mouse and Drosophila, respectively. Standard errors and significance of expressional differences, based on non-parametric Kolmogorov-Smirnov tests, between retrogenes and parental genes are indicated.

3.3.10 Analysis of sperm retrogene expression during spermatogenesis

Mammalian and Drosophila sperm retrogenes have different patterns of expression during spermatogenesis both when compared to each other and to their parental genes (Figure 3.5). In mammals, there is higher parental expression during early spermatogenesis, and higher sperm retrogenes expression at later stages. While in Drosophila sperm retrogenes are similarly expressed, relative to their parental genes, during early spermatogenesis, but higher during later stages.

3.3.11 Spermatogenesis expression of mammalian sperm retrogenes is dependent on the direction of retrotransposition event

Further analysis of mammalian sperm retrogenes revealed differences between sperm retrogene and their parent genes expression depending on the direction of retrogene movement (Figure 3.6). Mammalian sperm retrogenes that have moved between autosomes had significantly lower average expression during mitosis than their parental genes, but do not significantly differ during the later stages of spermatogenesis. Conversely, mammalian retrogenes, which have relocated from the X chromosome to an autosome, have significantly higher average expression during the later stages of spermatogenesis, but not mitosis, compared to their parental genes. This analysis was not repeated in Drosophila due to the small number of X to autosome transposition events with expression data.

3.3.12 Sperm retrogenes have distinct expression patterns compared to non-sperm retrogenes

We compared the expression profiles of sperm retrogenes to those of retrogenes that do not encode sperm components (termed non-sperm retrogenes). We observed that sperm retrogenes and non-sperm retrogenes were different in terms of levels of expression in the testis, during spermatogenesis and in their relation with their parental genes. Sperm retrogenes have higher expression in the testis compared to non-sperm retrogenes in both mammals and Drosophila (Figure 3.4). In addition, Drosophila sperm retrogenes demonstrate higher expression in all stages of spermatogenesis compared to non-sperm retrogenes, while mammalian sperm retrogenes are only higher later in spermatogenesis compared to non-sperm retrogenes (Figure 3.5). When we compared non-sperm retrogenes to their parent gene expression we found that, similar to mammalian sperm retrogenes, mammalian non-sperm retrogenes have similar levels of expression in the testis but much lower expression in somatic tissues compared to their parent genes. While, unlike Drosophila sperm retrogenes, which have higher expression in the testis compared to their parent genes, Drosophila non-sperm retrogenes have similar levels of testis expression to their parent genes (Figure 3.4). Finally, both mammalian and Drosophila sperm retrogenes have significantly higher expression during later spermatogenic stages compared to their parent genes, but non-sperm retrogenes do not (Figure 3.5).

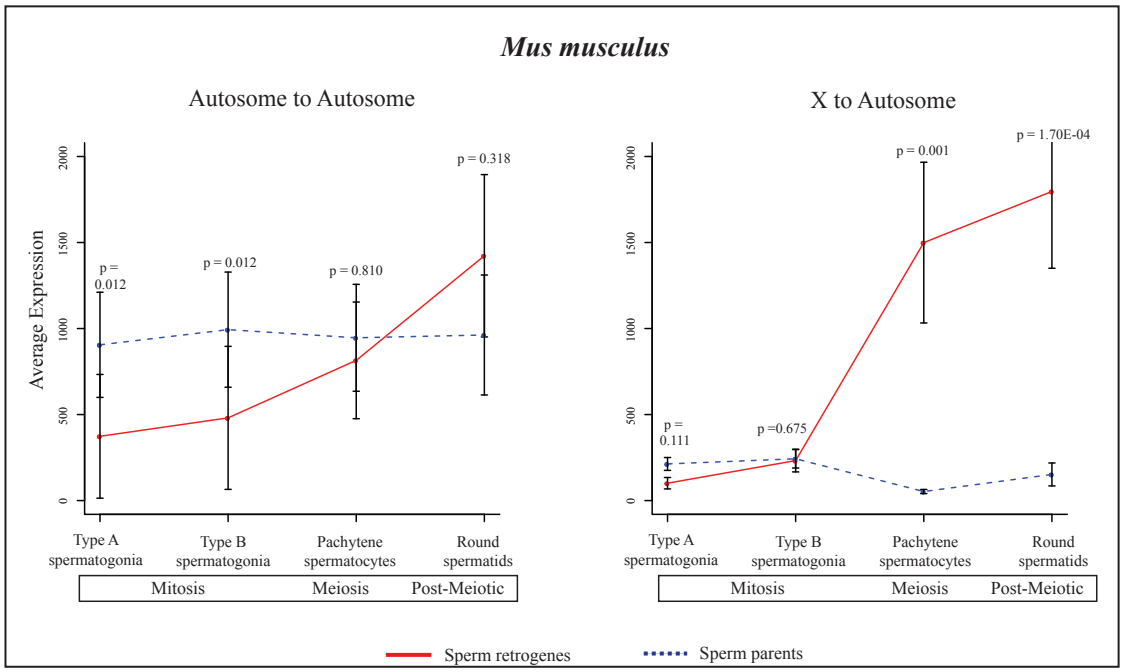

Figure 3.6 Mammalian sperm retrogene and parent gene expression during spermatogenesis. Average sperm retrogene (solid, red) and parental (dotted, blue) expression during the spermatogenic stages is displayed for mouse sperm retrogenes and their parents for (a) those that have relocated from autosome-to-autosome and (b) those that have relocated from the X chromosome to the autosomes. Expression data was obtained from Mammalian Reproductive Genetics (http://mrg.genetics.washington.edu). Standard errors and significance of expressional differences between average retrogene and parent expression, based on non-parametric Kolmogorov-Smirnov tests, are indicated.

3.4. Discussion

Consistent with proposals that gene duplication is an integral process in the creation of novel male reproductive genes, an integrative analysis of proteomic and genomic data revealed that ~20% of known retrogenes in both mammals and Drosophila encode sperm components identified by mass spectrometry (sperm retrogenes). Therefore it seems likely that retrotransposition is an important common mechanism for the creation of novel sperm genes in both mammals and Drosophila. However, it is likely that this number is an underestimate due to the conservative method of both retrogene and sperm protein identification. For example, the retrogene *k81* has not been identified in the *D. melanogaster* sperm proteome, despite its demonstrated function in sperm telomere maintenance (Loppin et al. 2005; Dubruille et al. 2010). It is noteworthy that in both mammals and Drosophila an enrichment of sperm retrogenes with functions in metabolism was observed. However, the proportion of sperm retrogenes with functions in metabolism (>70%) is significantly greater than would be predicted based on either the functional composition of the genomes or sperm proteomes (Tables 3.1 and 3.2). Similarly, the disparate enrichment of Drosophila and mammalian sperm retrogenes in different metabolic pathways (mitochondrial and carbohydrate metabolism, respectively) cannot be explained by underlying differences in the composition of either the genomes or sperm proteomes. Given the importance of metabolic processes in sperm function, and the role of spermatozoa in reproduction, it is likely that sperm metabolism has been the focus of intense sexual selection. As gene duplication is an integral mechanism in the generation of biological novelty we suggest that retrotransposition has been important in the evolution of novel sperm genes that enhance male reproductive fitness through roles in metabolism. Further, it is possible that the differential enrichment of metabolism-related sperm retrogenes in mammalian and Drosophila may be a reflection of the phylogenetic differences in the utilization or importance of these metabolic pathways in the spermatozoa of the two taxa. In the following sections we discuss how enrichment of sperm retrogenes has occurred in pathways that have been shown to be important to sperm fitness, and how these pathways differ between mammals and Drosophila.

3.4.1 Mammalian sperm retrogenes have functions related to energy provision for sperm motility and capacitation

Mammalian spermatozoa are generally relatively small highly motile cells, and sperm competitive ability appears to be related to sperm motility (Gage 1998; Gomendio and Roldan 2008; Firman and Simmons 2011). In addition, ATP for sperm motility is primarily generated by glycolysis via enzymes attached to the fibrous sheath and distributed along the length of the flagellum (Krisfalusi et al. 2006; Nascimento et al. 2008). Our analysis identified several sperm retrogenes with functions in glycolysis and related pathways, including previously characterised

retrogenes produced by parent genes that encode glycolytic enzymes (*Pgk2, Aldoart1*) (Vemuganti et al. 2007; Vemuganti, de Villena, and O'Brien 2010). In addition to sperm motility considerable energy is also required for sperm maturation, capacitation and the acrosome reaction. These processes may also be important in determining the outcome of sperm competition, and genes involved in these processes are also likely to be under sexual selection. As such it seems likely that gene duplication, of metabolic-related genes, may also enhance efficiency of these processes, and therefore sperm fitness. In support of the hypothesis that retrotransposition has enhanced sperm capacitation we identified mammalian sperm retrogenes in functions related to the control of these processes. Sperm capacitation, and the initiation of motility, is regulated in part by kinase activity and has been demonstrated in hamsters as dependent upon tyrosine phosphorylation of *Pdha2*, a mammalian sperm retrogene (Kumar, Rangaraj, and Shivaji 2006). We also observed a GO functional enrichment of mammalian sperm retrogenes functioning in phosphorylation, including *Gk2, Gykl1, Prps1l1* and two lineage specific kinases, the human *PRKACG* and the rat *Prkar1a*. In addition, two further mammalian sperm retrogenes, *G6pd2* and *PRPS1L1*, were identified in the pentose phosphate pathway, which is involved in human sperm capacitation (Miraglia et al. 2010). Thus the relative enhancement of selection upon sperm capacitation and motility in mammals may be responsible for the observed functional enrichment of sperm retrogenes involved in process that support these functions.

3.4.2 Drosophila sperm retrogenes have functions related to sperm development

In contrast to their mammalian counterparts, Drosophila spermatozoa are much larger, are produced in fewer numbers, and undergo a much longer period of sperm storage. As previously discussed, long sperm outcompete shorter ones in Drosophila (Miller and Pitnick 2002). This is not due to improved sperm motility, rather, longer sperms are more difficult to displace (Lüpold et al. 2012), suggesting that sperm length is under intense sexual selection similar to motility in mammals (Miller and Pitnick 2002). It has been proposed that the evolution of the nebenkern has been essential to the evolution of sperm gigantism, and therefore the morphological change that underlies the competitive advantage of long sperm (Noguchi, Koizumi, and Hayashi 2012). Although it is possible that residual ATP generated by the nebenkern enhances sperm motility, the role of sperm mitochondrial derivatives as energy producing organelles has been called into question (reviewed in Werner and Simmons 2008). However, an association between microtubule dynamics and nebenkern elongation has been demonstrated to be structurally essential for uniaxial spermatid tail development. With the disruption of the linkage between nebenkern and microtubules resulting in developmental defects of the flagellum and male sterility (Noguchi, Koizumi, and Hayashi 2011). We therefore suggest that the functional enrichment of sperm retrogenes in roles associated with mitochondrial energetics may be

associated with a role in the nebenkern, as a specialised organising centre for microtubules during sperm flagellum elongation. Sperm retrogenes may improve fitness by providing energy for both spermatogenesis in general and for the development of this specialised structure in particular. This is supported by the observation that Drosophila sperm retrogenes tend to be highly expressed during all stages of spermatogenesis (Figure 3.5). It is also noteworthy that several other sperm retrogenes have structural and sperm developmental functions, including: *Cdlc2*, which is involved in microtubule based movement and whose parent gene, *ctp*, is required for actin filament assembly in the elongating flagellum (Ghosh-roy, Desai, and Ray 2005); and *Act87E*, which is involved in cytoskeleton organisation. There are also sperm retrogenes with functions in sperm generation (*Gskt*) (Kalamegham et al. 2007), sperm development (*Hsp60B*) (Castrillon et al. 1993) and spermatid nucleus elongation and sperm individualisation (*Prosα6T*) (Zhong and Belote 2007). Further support is provided by the observation that the majority of the Drosophila sperm retrogenes originate prior to the diversification of the Drosophila genus and therefore predate the evolution of sperm gigantism in this taxon.

3.4.3 Comparison of sperm and non-sperm retrogenes

Sperm retrogenes are distinct from non-sperm retrogenes in both mammals and Drosophila, not only in their identification as encoding proteins found in mature spermatozoa but also in their patterns of expression. Consistent with previous studies on all retrogenes in mammals (Emerson et al. 2004) and Drosophila (Betrán, Thornton, and Long 2002), both sperm and non-sperm retrogenes have higher expression in the whole testis compared to somatic tissue. However, in both Drosophila and mammals, sperm retrogenes have higher expression in the testis than non-sperm retrogenes (Figure 3.4). Similarly, during spermatogenesis Drosophila sperm retrogenes have consistently higher expression at all stages than non-sperm retrogenes. It should be noted, however, that during mammalian spermatogenesis, sperm and non-sperm retrogenes have similar levels of expression during mitosis, and it is only during the later stages of spermatogenesis that sperm retrogenes dramatically increase their expression compared to non-sperm retrogenes (Figure 3.5). While the higher expression of sperm retrogenes in male reproductive tissues is consistent with the presence of their proteins in spermatozoa, non-sperm retrogenes are also expressed in these tissues, albeit at much lower levels. The expression of non-sperm retrogenes in the testis may be due to the whole testis samples containing both somatic and germline tissue. Similarly the expression of non-sperm retrogenes during spermatogenesis may be due to their expression early in spermatogenesis prior to the removal of cytoplasm and proteins from the developing spermatocyte. Alternatively, although less likely, these retrogenes may contribute proteins to the sperm proteome that have not yet been identified. Finally, the largest differences in expression between sperm retrogenes and both

parent genes and non-sperm retrogenes occur during the meiotic and post-meiotic stages of spermatogenesis in both mammals and Drosophila. This may be due to meiotic and post-meiotic stages of gametogenesis being the most specialised, distinct stages between males and females, and that these novel sperm genes have acquired functions during this "gender-specific" stage.

3.4.4 Comparison of the origination of sperm retrogenes in Drosophila and mammals

We observed a significantly higher proportion of sperm retrogenes with parental genes encoding sperm components in mammals than in Drosophila. This difference may occur from (1) biases in the identification of parental genes within the sperm proteome, (2) differences in the frequency of *de novo* sperm retrogene creation from non-sperm genes, or (3) differences in the extent to which sperm retrogenes acquire the functions of their progenitors. However, it is unlikely that that the differences are due to biases in the identification of parental genes within sperm proteomes. There are no reasons why mammalian sperm retrogene parents but not Drosophila sperm retrogene parents should be identified: mass spectrometry has identified similar numbers of proteins in the *D. melanogaster*, mouse, rat and human sperm proteomes, and there is substantial functional conservation between mouse and *D. melanogaster* sperm proteomes (Rettie and Dorus 2012).

It is difficult to determine whether the differences between the presence of sperm retrogenes parents in the sperm proteome is due to differences in *de novo* sperm gene creation or differences in the partitioning of ancestral gene function. A newly created gene may have several possible evolutionary fates: non-functionality, where the new gene becomes a pseudogene; neofunctionalization, where the new gene evolves a novel function; or subfunctionalization, where the gene duplicate acquires some or all functions of the original gene (Lynch and Katju 2004; Kaessmann 2010). However, it is impossible to determine whether there has been a bias in *de novo* creation of novel sperm genes without detailed information on either multiple closely related species or the ancestral sperm proteome. Without knowing whether or not the parents of sperm retrogenes are/were present in any sperm proteome, we can only speculate on whether there has been partial subfunctionalization (e.g. spatially, between tissues, or temporally) or neofunctionalization/complete subfunctionalization of the ancestral gene function using expression data. In mammals we can infer that there has been a temporal division of functions between the parent and retrogene: both have similar levels of expression in the testis, many parental genes encode sperm proteins themselves, and there is a distinct decrease in parental gene express during meiosis with a corresponding increase in sperm retrogene expression during this phase. While in Drosophila it is possible that there has been either complete spatial separation of function or neofunctionalization, as the sperm retrogenes

have significantly higher expression in both the testis and spermatogenesis compared to their parental genes. Although this is not direct evidence for subfunctionalization, there are compelling arguments that also support these inferences. In mammals, the idea of temporal subfunctionality has previously been proposed due to the existence of meiotic sex chromosome inactivation (MSCI; discussed below), while in Drosophila the creation of gene duplicates from nuclear mitochondrial genes has been hypothesized to be associated with the resolution of intralocus conflict. Gallach *et al.* (2010) observed that a large number of gene duplicates originated from nuclear mitochondrial genes. It was proposed that these duplicates resolved the genomic antagonism caused by the increased energy needs of the testis that had the potential to be very damaging if expressed in the somatic tissue. Such excess in energy is unnecessary in somatic cells with much lower energy requirements. The creation of a testis specific retrogene allows high levels of mitochondrial derived energy in the testis but not in somatic tissue where the by-product of mitochondrial metabolism (reactive oxygen species) could reduce the overall fitness of the organism (Gallach, Chandrasekaran, and Betrán 2010). However, while not precluding this explanation, the potential that gene duplication has provided genetic novelty that has enhanced male fitness may be more likely to be behind the observed functional enrichment of sperm retrogenes. Furthermore, if the avoidance of increased mitochondrial activity in somatic tissues has selected for testis-specific gene duplications, why do we not observe a similar number of mammalian sperm retrogenes with functions in mitochondrial derived energetics, an organelle that is also highly abundant in mammalian sperm?

In both mammals and Drosophila, the X chromosome has been observed to produce an excess of retrogenes that relocate to the autosomes (Betrán, Thornton, and Long 2002; Emerson et al. 2004) with male-biased genes being under-represented (Parisi et al. 2003; Khil et al. 2004; Sturgill et al. 2007). This under-representation of genes on the X chromosome has been attributed to various factors including avoidance of sexual antagonism and meiotic sex chromosome inactivation (MSCI) (Betrán, Thornton, and Long 2002; Wu and Xu 2003). In relation to X chromosome linkage it is theorised that male-biased genes are more likely to reside on the autosomes due to the relative amount of (evolutionary) time the X chromosome spends in males compared to females (Wu and Xu 2003). Similarly, the existence of MSCI, where the sex chromosomes are transcriptionally silenced during the meiotic phase of spermatogenesis, affects X-linked gene content, as the X chromosome is not an advantageous location for any gene required during meiosis. For example, while the mouse X chromosome is generally under-represented for testis-expressed genes this is not observed for genes expressed in the mitotic or post-meiotic stages of spermatogenesis (Khil et al. 2004; Mueller et al. 2008). Finally, it should be noted that while MSCI is a well-established phenomenon in mammals it remains controversial in Drosophila (Hense, Baines, and Parsch 2007; Vibranovski et al. 2009;

Meiklejohn et al. 2011; Mikhaylova and Nurminsky 2011). Consistent with the X chromosome as a disadvantageous location for male genes, we observed no X-linked sperm retrogenes in either mammals or Drosophila. However, we observed that there has been an excess of mammalian, but not Drosophila, sperm retrogenes moving from the X chromosome to the autosomes. This difference is observed when we considered the genomic distribution of all genes, sperm proteome genes, metabolic genes and sperm proteome genes with functions in metabolism. Therefore, we propose that there is evidence that MSCI is an important factor in the evolution of novel sperm genes in mammals, but not necessarily in Drosophila, based on (1) the significant excess of retrotransposition events off the X chromosome, (2) the number of X-linked parental genes that also encode sperm proteins (7 out of 10), and (3) the complementary expression patterns of sperm retrogenes and their parents during spermatogenesis that is primarily driven by X-to-autosome retrotransposition events (Figure 3.6). While the complementary pattern of parent and retrogene expression during spermatogenesis does not confirm compensatory expression of the X-linked parental gene by the autosomal retrogene, one of the X-to-autosome sperm retrogenes with a parental gene in sperm, *Pgk2*, has been shown to compensate for its parental gene inactivation during spermatogenesis (Danshina et al. 2010). Together this suggests that MSCI has been particularly influential in the evolution of mammalian sperm. However, Drosophila sperm retrogenes do not show an excess of X to autosome movements, nor do they have parental genes currently identified in the *D. melanogaster* sperm proteome and their expression during spermatogenesis does not infer a complementary pattern of expression with their parental genes. While this does not support or reject the existence of MSCI in Drosophila, it does suggest that MSCI has not been influential in the evolution of Drosophila sperm retrogenes.

3.4.4 Summary

Spermatozoa are essential for male reproduction, and all spermatozoa perform the shared function of delivering the paternal genetic contribution. Due to their importance in reproduction, spermatozoa are under intense sexual selection mediated by sperm competition, that is responsible for the morphological diversity of sperm (Pitnick, Hosken, and Birkhead 2009; Gage 2012). Gene duplication has contributed to the creation of novel sperm genes that encode sperm components and have enhanced sperm competitive ability (e.g. *Sdic* and *tektin*). In this study we have demonstrated that retrotransposition has significantly contributed to the sperm proteome composition in both mammals and Drosophila. However, we also observed that while retrogenes in both lineages function in sperm metabolism, an important process in sperm fitness, they have functions in different metabolic pathways and these pathways mirror the disparate processes under selection in mammals (motility and capacitation) and Drosophila (sperm length). Therefore, we propose that (1) novel Drosophila sperm retrogenes were

retained in the genome with mitochondrial energetic functions due to the increasing demand to provide energy for the elongation of an ever increasing flagellum length, due to co-evolution with the female reproductive tract, and (2) novel mammalian sperm retrogenes were retained in carbohydrate energetic functions to provide additional glycolytic enzymes for the substantial energy demands of sperm motility and acquisition of fertilization competence. Ultimately, targeted investigations will be required to establish the specific functions of these novel sperm genes and infer the differing selective forces associated with their evolutionary retention.

3.5 Acknowledgments

I would like to thank Dr. Catherine Pink for creating a computer script to facilitate the processing of the mammalian BLAST results.

3.6 References

Anderson, M. J., J. Nyholt, and A. F. Dixson. 2005. "Sperm Competition and the Evolution of Sperm Midpiece Volume in Mammals." *Journal of Zoology* 267 (2): 135–142.

Bai, Yongsheng, Claudio Casola, Cédric Feschotte, and Esther Betrán. 2007. "Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in Drosophila." *Genome Biology* 8 (1): R11.

Baker, Mark A, Louise Hetherington, Gabi M Reeves, and R John Aitken. 2008. "The Mouse Sperm Proteome Characterized via IPG Strip Prefractionation and LC-MS/MS Identification." *Proteomics* 8 (8): 1720–1730.

Baker, Mark A, Louise Hetherington, Gabi Reeves, Jörg Müller, and R John Aitken. 2008. "The Rat Sperm Proteome Characterized via IPG Strip Prefractionation and LC-MS/MS Identification." *Proteomics* 8 (11): 2312–2321.

Baker, Mark A, Gabi Reeves, Louise Hetherington, Jörg Müller, Inke Baur, and R John Aitken. 2007. "Identification of Gene Products Present in Triton X-100 Soluble and Insoluble Fractions of Human Spermatozoa Lysates Using LC-MS/MS Analysis." *Proteomics. Clinical Applications* 1 (5): 524–532.

Belote, J M, and L Zhong. 2009. "Duplicated Proteasome Subunit Genes in Drosophila and Their Roles in Spermatogenesis." *Heredity* 103 (1): 23–31.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. "Retroposed New Genes Out of the X in Drosophila." *Genome Research* 12 (12): 1854–1859.

Bradley, Julie, Andrew Baltus, Helen Skaletsky, Morgan Royce-Tolland, Ken Dewar, and David C Page. 2004. "An X-to-autosome Retrogene Is Required for Spermatogenesis in Mice." *Nature Genetics* 36 (8): 872–976.

Cao, Wenlei, George L Gerton, and Stuart B Moss. 2006. "Proteomic Profiling of Accessory Structures from the Mouse Sperm Flagellum." *Molecular & Cellular Proteomics* 5 (5): 801–810.

Castrillon, Diego H, Pierre Gonczy, Sherry Alexander, Robert Rawson, Charles G Eberhart, Sridhar Viswanathan, Stephen DiNardo, and Steven A Wasserman. 1993. "Toward a Molecular Genetic Analysis of Spermatogenesis in *Drosophila melanogaster*: Characterization of Male-sterile Mutants Generated by Single P Element Mutagenesis." *Genetics* 135 (2): 489–505.

Chintapalli, Venkateswara R, Jing Wang, and Julian A T Dow. 2007. "Using FlyAtlas to Identify Better *Drosophila melanogaster* Models of Human Disease." *Nature Genetics* 39 (6): 715–720.

Danshina, Polina V, Christopher B Geyer, Qunsheng Dai, Eugenia H Goulding, William D Willis, G Barrie Kitto, John R McCarrey, E M Eddy, and Deborah A O'Brien. 2010. "Phosphoglycerate Kinase 2 (PGK2) Is Essential for Sperm Function and Male Fertility in Mice." *Biology of Reproduction* 82 (1): 136–145.

Dorus, Steve, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L Karr. 2006. "Genomic and Functional Evolution of the Drosophila Melanogaster Sperm Proteome." *Nature Genetics* 38 (12): 1440–1445.

Dorus, Steve, Zoë N Freeman, Elizabeth R Parker, Benjamin D Heath, and Timothy L Karr. 2008. "Recent Origins of Sperm Genes in Drosophila." *Molecular Biology and Evolution* 25 (10): 2157–2166.

Dorus, Steve, Elizabeth R Wasbrough, Jennifer Busby, Elaine C Wilkin, and Timothy L Karr. 2010. "Sperm Proteomics Reveals Intensified Selection on Mouse Sperm Membrane and Acrosome Genes." *Molecular Biology and Evolution* 27 (6): 1235–1246.

Dorus, Steve, Elaine C Wilkin, and Timothy L Karr. 2011. "Expansion and Functional Diversification of a Leucyl Aminopeptidase Family That Encodes the Major Protein Constituents of Drosophila Sperm." *BMC Genomics* 12: 177.

Dubruille, Raphaëlle, Guillermo A Orsi, Lætitia Delabaere, Elisabeth Cortier, Pierre Couble, Gabriel A B Marais, and Benjamin Loppin. 2010. "Specialization of a Drosophila Capping Protein Essential for the Protection of Sperm Telomeres." *Current Biology* 20 (23): 2090–2099.

Emerson, J J, Henrik Kaessmann, Esther Betrán, and Manyuan Long. 2004. "Extensive Gene Traffic on the Mammalian X Chromosome." *Science* 303 (5657): 537–540.

Firman, Renée C, and Leigh W Simmons. 2011. "Experimental Evolution of Sperm Competitiveness in a Mammal." *BMC Evolutionary Biology* 11: 19.

Gage, M J. 1998. "Mammalian Sperm Morphometry." *Proceedings of the Royal Society B - Biological Sciences* 265 (1391): 97–103.

Gage, Matthew J.G., Christopher P. Macfarlane, Sarah Yeates, Richard G. Ward, Jeremy B. Searle, and Geoffrey A. Parker. 2004. "Spermatozoal Traits and Sperm Competition in Atlantic Salmon: Relative Sperm Velocity Is the Primary Determinant of Fertilization Success." *Current Biology* 14 (1): 44–47.

Gage, Matthew J G. 2012. “Complex Sperm Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (12): 4341–4342.

Gallach, Miguel, Chitra Chandrasekaran, and Esther Betrán. 2010. “Analyses of Nuclearly Encoded Mitochondrial Genes Suggest Gene Duplication as a Mechanism for Resolving Intralocus Sexually Antagonistic Conflict in Drosophila.” *Genome Biology and Evolution* 2: 835–850.

Ghosh-roy, Anindya, Bela S Desai, and Krishanu Ray. 2005. “Dynein Light Chain 1 Regulates Dynamin-mediated F-Actin Assembly During Sperm Individualization in Drosophila.” *Molecular Biology of the Cell* 16 (7): 3107–3116.

Gibbons, R, S A Adeoya-Osiguwa, and L R Fraser. 2005. “A Mouse Sperm Decapacitation Factor Receptor Is Phosphatidylethanolamine-binding Protein 1.” *Reproduction* 130 (4): 497–508.

Gomendio, Montserrat, and Eduardo R S Roldan. 2008. “Implications of Diversity in Sperm Size and Function for Sperm Competition and Fertility.” *The International Journal of Developmental Biology* 52 (5-6): 439–447.

Greenspan, Leah, and Andrew G Clark. 2011. “Associations Between Variation in X Chromosome Male Reproductive Genes and Sperm Competitive Ability in *Drosophila melanogaster*.” *International Journal of Evolutionary Biology* 2011: 214280–214289.

Hense, Winfried, John F Baines, and John Parsch. 2007. “X Chromosome Inactivation During Drosophila Spermatogenesis.” *PLoS Biology* 5 (10): 2288–2295.

Hosken, David J, and Paul I Ward. 2001. “Experimental Evidence for Testis Size Evolution via Sperm Competition.” *Ecology Letters* 4 (1): 10–13.

Kaessmann, Henrik. 2010. “Origins, Evolution, and Phenotypic Impact of New Genes.” *Genome Research* 20 (10): 1313–1326.

Kalamegham, Rasika, David Sturgill, Esther Siegfried, and Brian Oliver. 2007. “Drosophila Mojoless, a Retroposed GSK-3, Has Functionally Diverged to Acquire an Essential Role in Male Fertility.” *Molecular Biology and Evolution* 24 (3): 732–742.

Karr, Timothy L, and Scott Pitnick. 1996. “The Ins and Outs of Fertilization.” *Nature* 379 (6564): 405–406.

Khil, Pavel P, Natalya A Smirnova, Peter J Romanienko, and R Daniel Camerini-Otero. 2004. “The Mouse X Chromosome Is Enriched for Sex-biased Genes Not Subject to Selection by Meiotic Sex Chromosome Inactivation.” *Nature Genetics* 36 (6): 642–646.

Kleven, Oddmund, Terje Laskemoen, Frode Fossøy, Raleigh J Robertson, and Jan T Lifjeld. 2008. “Intraspecific Variation in Sperm Length Is Negatively Related to Sperm Competition in Passerine Birds.” *Evolution* 62 (2): 494–499.

Krisfalusi, Michelle, Kiyoshi Miki, Patricia L Magyar, and Deborah A O’Brien. 2006. “Multiple Glycolytic Enzymes Are Tightly Bound to the Fibrous Sheath of Mouse Spermatozoa.” *Biology of Reproduction* 75 (2): 270–278.

Kumar, Vivek, Nandini Rangaraj, and Sisinthy Shivaji. 2006. “Activity of Pyruvate Dehydrogenase A (PDHA) in Hamster Spermatozoa Correlates Positively with Hyperactivation and Is Associated with Sperm Capacitation.” *Biology of Reproduction* 75 (5): 767–777.

LaMunyon, Craig W, and Samuel Ward. 2002. “Evolution of Larger Sperm in Response to Experimentally Increased Sperm Competition in *Caenorhabditis elegans*.” *Proceedings of the Royal Society B - Biological Sciences* 269 (1496): 1125–1128.

Loppin, Benjamin, David Lepetit, Steve Dorus, Pierre Couble, and Timothy L Karr. 2005. “Origin and Neofunctionalization of a Drosophila Paternal Effect Gene Essential for Zygote Viability.” *Current Biology* 15 (2): 87–93.

Lynch, Michael, and John S Conery. 2003. “The Origins of Genome Complexity.” *Science* 302 (5649): 1401–1404.

Lynch, Michael, and Vaishali Katju. 2004. “The Altered Evolutionary Trajectories of Gene Duplicates.” *Trends in Genetics* 20 (11): 544–549.

Lüpold, Stefan, Mollie K Manier, Kirstin S Berben, Kyle J Smith, Bryan D Daley, Shannon H Buckley, John M Belote, and Scott Pitnick. 2012. “How Multivariate Ejaculate Traits Determine Competitive Fertilization Success in *Drosophila melanogaster*.” *Current Biology* 22 (18): 1667–1672.

Malo, Aurelio F, J Julián Garde, Ana J Soler, Andrés J García, Montserrat Gomendio, and Eduardo R S Roldan. 2005. “Male Fertility in Natural Populations of Red Deer Is Determined by Sperm Velocity and the Proportion of Normal Spermatozoa.” *Biology of Reproduction* 72 (4): 822–829.

Meiklejohn, Colin D, Emily L Landeen, Jodi M Cook, Sarah B Kingan, and Daven C Presgraves. 2011. “Sex Chromosome-specific Regulation in the Drosophila Male Germline but Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation.” *PLoS Biology* 9 (8): e1001126.

Mikhaylova, Lyudmila M, and Dmitry I Nurminsky. 2011. “Lack of Global Meiotic Sex Chromosome Inactivation, and Paucity of Tissue-specific Gene Expression on the Drosophila X Chromosome.” *BMC Biology* 9: 29.

Miller, Gary T, and Scott Pitnick. 2002. “Sperm-female Coevolution in Drosophila.” *Science* 298 (5596): 1230–1233.

Miraglia, E, C Lussiana, D Viarisio, C Racca, A Cipriani, E Gazzano, A Bosia, A Revelli, and D Ghigo. 2010. “The Pentose Phosphate Pathway Plays an Essential Role in Supporting Human Sperm Capacitation.” *Fertility and Sterility* 93 (7): 2437–2440.

Morrow, Edward H, and Matthew J G Gage. 2001. “Consistent Significant Variation Between Individual Males in Spermatozoal Morphometry.” *Journal of Zoology* 254: 147–153.

Mueller, Jacob L, Shantha K Mahadevaiah, Peter J Park, Peter E Warburton, David C Page, and James M A Turner. 2008. “The Mouse X Chromosome Is Enriched for Multicopy Testis Genes Showing Postmeiotic Expression.” *Nature Genetics* 40 (6): 794–799.

Nascimento, Jaclyn M, Linda Z Shi, James Tam, Charlie Chandsawangbhuwana, Barbara Durrant, Elliot L Botvinick, and Michael W Berns. 2008. “Comparison of Glycolysis and Oxidative Phosphorylation as Energy Sources for Mammalian Sperm Motility, Using the Combination of Fluorescence Imaging, Laser Tweezers, and Real-time Automated Tracking and Trapping.” *Journal of Cellular Physiology* 217 (3): 745–751.

Noguchi, Tatsuhiko, Michiko Koizumi, and Shigeo Hayashi. 2011. “Sustained Elongation of Sperm Tail Promoted by Local Remodeling of Giant Mitochondria in Drosophila.” *Current Biology* 21 (10): 805–814.

Noguchi, Tatsuhiko, Michiko Koizumi, and Shigeo Hayashi. 2012. “Mitochondrai-driven Cell Elongation Mechanism for Competing Sperms.” *Fly* 6 (2): 113–116.

Okuda, Hidenobu, Akira Tsujimura, Shinji Irie, Keisuke Yamamoto, Shinichiro Fukuhara, Yasuhiro Matsuoka, Tetsuya Takao, et al. 2012. “A Single Nucleotide Polymorphism Within the Novel Sex-linked Testis-specific Retrotransposed PGAM4 Gene Influences Human Male Fertility.” *PloS One* 7 (5): e35195.

Parisi, Michael, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. 2003. “Paucity of Genes on the Drosophila X Chromosome Showing Male-biased Expression.” *Science* 299 (5607): 697–700.

Parker, GA. 1970. “Sperm Competition and Its Evolutionary Consequences in Insects.” *Biol Rev Camb Philos Soc* 45: 525–67.

Pitnick, S, DJ Hosken, and TR Birkhead. 2009. “Sperm Morphological Diversity.” In *Sperm Biology: An Evolutionary Perspective*, ed. Pitnick S Birkhead TR, Hosken DJ, 69–149. USA: Academic Press.

Rautureau, Gilles, Laurence Jouvensal, Françoise Vovelle, Françoise Schoentgen, Daniel Locker, and Martine Decoville. 2009. “Expression and Characterization of the PEBP Homolog Genes from Drosophila.” *Archives of Insect Biochemistry and Physiology* 71 (2): 55–69.

Rettie, Elaine C, and Steve Dorus. 2012. “Drosophila Sperm Proteome Evolution - Insights from Comparative Genomic Approaches.” *Spermatogenesis* 2 (3): 213–223.

Stein, Kathryn K, Jowell C Go, William S Lane, Paul Primakoff, and Diana G Myles. 2006. “Proteomic Analysis of Sperm Regions That Mediate Sperm-egg Interactions.” *Proteomics* 6 (12): 3533–3543.

Sturgill, David, Yu Zhang, Michael Parisi, and Brian Oliver. 2007. “Demasculinization of X Chromosomes in the Drosophila Genus.” *Nature* 450 (7167): 238–242.

Vemuganti, Soumya A, Timothy A Bell, Cameron O Scarlett, Carol E Parker, Fernando Pardo-Manuel de Villena, and Deborah A O’Brien. 2007. “Three Male Germline-specific Aldolase A Isozymes Are Generated by Alternative Splicing and Retrotransposition.” *Developmental Biology* 309 (1): 18–31.

Vemuganti, Soumya A, Fernando Pardo-Manuel de Villena, and Deborah A O’Brien. 2010. “Frequent and Recent Retrotransposition of Orthologous Genes Plays a Role in the Evolution of Sperm Glycolytic Enzymes.” *BMC Genomics* 11: 285.

Vibranovski, Maria D, Hedibert F Lopes, Timothy L Karr, and Manyuan Long. 2009. “Stage-specific Expression Profiling of Drosophila Spermatogenesis Suggests That Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-expressed Genes.” *PLoS Genetics* 5 (11): e1000731.

Wasbrough, Elizabeth R, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L Karr. 2010. “The *Drosophila melanogaster* Sperm proteome-II (DmSP-II).” *Journal of Proteomics* 73 (11): 2171–2185.

Werner, Michael, and Leigh W Simmons. 2008. “Insect Sperm Motility.” *Biological Reviews* 83 (2): 191–208.

Wu, Chung-I, and Eugene Yujun Xu. 2003. “Sexual Antagonism and X Inactivation – the SAXI Hypothesis.” *Trends in Genetics* 19 (5): 243–247.

Yeh, Shu-Dan, Tiffanie Do, Carolus Chan, Adriana Cordova, Francisco Carranza, Eugene A Yamamoto, Mashya Abbassi, et al. 2012. “Functional Evidence That a Recently Evolved Drosophila Sperm-specific Gene Boosts Sperm Competition.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (6): 2043–2048.

Zhang, Ning, Junbo Liang, Yongqiang Tian, Ligang Yuan, Lan Wu, Shiying Miao, Shudong Zong, and Linfang Wang. 2010. “A Novel Testis-specific GTPase Serves as a Link to Proteasome Biogenesis : Functional Characterization of RhoS / RSA-14-44 in Spermatogenesis.” *Molecular Biology of the Cell* 21 (24): 4312–4324.

Zhong, Lei, and John M Belote. 2007. “The Testis-specific Proteasome Subunit Prosalpha6T of D. Melanogaster Is Required for Individualization and Nuclear Maturation During Spermatogenesis.” *Development* 134 (19): 3517–3525.

Zhou, Qi, Guojie Zhang, Yue Zhang, Shiyu Xu, Ruoping Zhao, Zubing Zhan, Xin Li, Yun Ding, Shuang Yang, and Wen Wang. 2008. “On the Origin of New Genes in Drosophila.” *Genome Research* 18 (9): 1446–1455.

Chapter 4

Genomic organisation of co-expressed genes: application of new analytical approach to testis expressed genes

4.1 Introduction

Gene order is non-random in both prokaryotes and eukaryotes. In prokaryotes genes are organised into operons: tightly packed, co-regulated and functionally related genes which have been evolutionarily conserved in Eubacteria and Archaea (Tamames 2001). Similarly, there is increasing evidence that eukaryotic genomes also exhibit non-random organization with respect to gene location (reviewed by Hurst, Pál, and Lercher 2004). This organisation can be observed both at the level of the chromosome, with non-random distributions of genes between chromosomes (particularly between autosomes and sex chromosomes) and also in terms of gene order along chromosomes. At the inter-chromosome level, significant differences in sex-biased gene expression has been demonstrated between autosomal and X-linked genes in Drosophila (Parisi et al. 2003; Sturgill et al. 2007), mouse (Khil et al. 2004; Mueller et al. 2008) and human (Lercher, Urrutia, and Hurst 2003), as well as between the Z chromosome and autosomes in chicken (Kaiser and Ellegren 2006). Within chromosomes, non-random spatially co-localised ("clustering") of co-expressed genes has been in a range of species, including yeast (Lercher and Hurst 2006), human (Lercher, Urrutia, and Hurst 2002), Drosophila (Boutanaev et al. 2002) and others (Ng, Wu, and Zhang 2009; Woo, Walker, and Churchill 2010;). Although the evolutionary conservation of gene order has been intensively studied in some cases, such as the HOX gene cluster which are responsible for the correct patterning of the embryos along the anterior-posterior axis (reviewed in Garcia-Fernàndez 2005), there are relatively few genome-level studies of gene cluster conservation. This chapter focuses on the development and application of a new analytical approach to characterize gene clustering upon genes co-expressed in the *Drosophila melanogaster* testis.

In eukaryotes the non-random organisation of genes that share temporal or spatial expression into spatially co-localised gene neighbourhoods has lead to the proposal that gene positional organisation is associated with transcriptional regulation (Schneider and Grosschedl 2007). While local regulation by *cis*-acting promoter sequences generally mediates individual gene expression (although examples of promoter sharing have been characterized; (Loppin et al. 2005), chromatin structure has been demonstrated to influence gene expression across larger regions containing numerous adjacent genes. The mechanism of chromatin based regulation is

largely associated with the accessibility of DNA to transcription factors (Cairns 2009), with changes in chromatin structure, resulting in "open" or "closed" conformations, generally enacted by post-translation modifications (e.g. acetylation or methylation) of histones or other chromatin associated proteins (Lelli, Slattery, and Mann 2012). Chromatin mediated co-regulation of adjacent genes would therefore logically explain the observed non-random distribution of co-expressed genes and their evolutionary conservation between species. Tremendous advances have been made in the empirical characterization of genome-wide chromatin structure through the application of chromatin immunoprecipitation (ChIP) techniques coupled with either with microarray tilling arrays (ChIP-chip) or RNA-sequencing (ChIP-seq) (Celniker et al. 2009). Despite our increased understanding of chromatin across a range of developmental stages and cell types, a detailed understanding of the relationship between chromatin domains, gene co-localisation and the spatial conservation of co-localised genes has yet to be attained. This is, in part, due to inconsistencies in the analytical approaches that have been used to characterize co-localized genes. In order for progress to be made in these enquires it is therefore necessary to develop bioinformatic tools that can reliably identify gene neighbourhoods from a range of species and using a range of different experimental data.

4.1.1 Methods for assessing genome organisation

The availability of high quality annotated genomes, in conjunction with expression, proteomic and gene ontology datasets, has made it bioinformatically possible to characterize the co-localisation genes based on a variety of gene characteristics. Several methods have been developed to assess the non-random clustering of genes and these can be largely divided into two main types: (1) methods that use a single focal gene to build neighbourhoods of genes with shared characteristics and (2) methods that assess evidence of shared gene characteristics within assigned genomic regions (defined either by physical distance or numbers of adjacent genes). Although fundamentally different, the patterns obtained using these approaches have been statistically assessed using random simulations in order to approximate stochastic genome organisation. In the following section, we discuss the assumptions associated with each approach and the primary observations obtained through their utilization.

One approach to assessing gene organisation is to begin with a single gene as a seed, for the assessment of a specified genomic region centred around this gene. For example, in *Saccharomyces cerevisiae* a fixed region consisting of an essential gene and 4 to 8 genes on either side (depending on analysis) were examined, and essential gene neighbourhoods were those in which more than 50% of gene in this fixed region were also essential genes (Batada and Hurst 2007). This method demonstrated that nucleosome occupancy was approximately half in essential gene neighbourhoods compared to other genomic regions, suggesting that these

essential gene neighbourhoods were within open chromatin regions (Batada and Hurst 2007). Despite this observation, this model has some drawbacks as it intrinsically assumes that the focal gene is in the centre of the gene neighbourhood under investigation, does not consider the distribution of genes within this neighbourhood. It also suffers from the same drawbacks of "sliding window" approaches (discussed below) as it arbitrarily parameterizes the size of gene neighbourhoods being analyzed. An alternative to this method, which does not make assumptions about neighbourhood size, begins with an individual gene from which an extended neighbourhood is constructed by the addition of close or neighbouring genes. These methods do not impose an upper limit on neighbourhood size, however they tend to include parameters that limit the range of candidate genes for addition to a gene neighbourhood. In *D. melanogaster* an estimate of co-expression (Pearson's correlation co-efficient, *r*) for all gene pairs in the genome was calculated, and random simulations determined the value of *r* at which genes were proposed to be significantly co-expressed. A pair of genes was then determined to be within a cluster if *r* was significant and no more than three intervening genes separated them. The cluster was then built by expanding to a neighbouring gene, which was no more than three intervening genes distant, with which the one of the existing gene pair and the new gene were significantly co-expressed (Weber and Hurst 2011). While this method is much more flexible than the first, it does not ensure that in the resulting gene cluster the average co-efficient for all genes is above the significance threshold of co-expression, rather that there is a correlation between neighbour proximity and co-expression. It also suffers from the need to assign an arbitrary value for the number of intervening genes. As selection for gene co-localisation may operate at a larger scale within chromatin domains or between chromatin insulators methods that assay expression within specified regions may be more effective.

In contrast to methods that begin with a single gene, other approaches systematically search the genome for regions that conform to specified parameters. One method is to divide the genome into equal, non-overlapping sections and determine the frequency of genes expressed in a certain tissue, process, cell type or expressed over a similar breadth of tissues. While simple, this method can be effective. For example, the genome of *Caenorhabditis elegans* was divided into regions of 0.5 recombination units, three of these regions contained significantly more genes expressed during spermatogenesis that expected (Miller et al. 2004). However, this method arbitrarily determines the boundaries and size of potential neighbourhoods and does not account for the possibility that biological meaningful domains may span the predetermined boundaries. An alternative approach is to employ a "sliding-window" method. The sliding window approach uses a pre-determined window size, usually a set number of genes or genomic distance, which is then used to sequentially survey the genome by moving along the chromosome in a directional manner by a succession of "steps", again usually a set number of

genes or genomic distance, and generally sequential windows overlap. At each step the region within the window is surveyed to determine whether it meets the requirements of a gene neighbourhood. These requirements may be that all, or a minimum proportion, of genes within the window are expressed in a given tissue (Boutanaev et al. 2002), or that the average index of co-expression within the window is greater than a specified threshold (Lercher and Hurst 2006; Sémon and Duret 2006). Overlapping or adjacent windows, in which the conditions for a neighbourhood have been meet, are then merged. Further conditions may also be added including: a maximum number of genes within the region that do not meet the requirements or a maximum distance between two adjacent genes (Li, Lee, and Zhang 2005). The consequence of this approach is that boundary locations are not arbitrarily determined. However, this method necessarily includes an arbitrary window size parameter and because over-lapping windows are coalesced this may lead to the identification of larger neighbourhoods than exist. Lastly, sliding window analyses present statistical complications because windows are overlapping and therefore lack independence. Although measures have been employed to account for this statistical issue it would be favourable to develop a model that does not suffer from this issue or unnecessarily set window size parameters that may detract from biological meaningful observations.

4.1.2 Characteristics of gene neighbourhoods

Non-random co-localisation of genes with similar expression has been observed in a diverse number of eukaryotic genomes (reviewed in Hurst, Pál, and Lercher 2004; Oliver and Misteli 2005; Michalak 2008; Koonin 2009). Many studies have focused on the organisation of ubiquitously expressed genes. These genes are expressed in all, or most, tissues, they tend to be associated with functions essential to the cell and are often termed housekeeping or essential genes due to their importance in maintaining the cell. In humans, housekeeping genes are significantly under-dispersed, and it has been proposed that there may be an advantage in the co-localisation of these genes within chromatin regions that are continuously open in all cells due to the need for them to be accessible in all cells (Lercher, Urrutia, and Hurst 2002). Further, these genes are thought to be especially sensitive to stochastic fluctuations in both the timing and amount of their expression. As such they are hypothesised to co-localise to regions of reduced random gene expression such as continuously open chromatin domains (Batada and Hurst 2007). While ubiquitously expressed genes are interesting they represent a small sub-set of the genome. For example, the human genome contains ~20,000-25,000 genes (International Human Genome Sequencing Consortium 2004), but only ~1,000 are expressed in all tissues, and of those <400 were considered to have high enough quality DNA sequences and RNA expression data to be used in the characterisation of the minimum human housekeeping transcriptome (Chiaromonte, Miller, and Bouhassira 2003). Other studies have focused on the

spatial organisation of genes that are similarly expressed across different tissues, time points or experimental conditions. Studies have demonstrated that neighbouring gene pairs tend to have higher measures of co-expression than gene pairs further apart in a wide variety of taxa, including yeast (Lercher and Hurst 2006), Drosophila (Spellman and Rubin 2002), zebra fish (Ng, Wu, and Zhang 2009) and mammals (Woo, Walker, and Churchill 2010). However, this effect extends beyond neighbouring gene pairs. In yeast regions of co-expressed genes have been observed spanning up to 30 genes (Lercher and Hurst 2006). Similarly in Drosophila 20% of the genome are co-localised into co-expressed regions that span 10 to 30 genes (Spellman and Rubin 2002). Further, similar to the evolutionary conservation of prokaryotic operons, orthologous neighbourhoods of co-expressed genes have been maintained since the divergence of the mouse and human genomes (Sémon and Duret 2006).

Lastly, a variety of studies have analyzed the co-localization of genes with tissue specific expression. It has been suggested that tissue-specific genes might be co-localised in order to be effectively silenced in tissues where they are not required, thereby reducing transcription noise and avoiding associated deleterious effects, while increasing the efficiency of their availability in the tissue in which they are needed (Shevelyov et al. 2009). This idea is supported as chromosomes themselves occupy distinct tissue specific spatial arrangements and as different regions of the nucleosome have distinct concentrations of transcription factors this may be related to tissue specific organisation for gene expression (Parada, McQueen, and Misteli 2004). Although support for this was lacking in humans when all tissues were considered (Lercher, Urrutia, and Hurst 2002), several studies have identified highly significant clustering of male-biased genes. Indeed non-random distribution of spermatogenesis expressed genes have been observed in yeast (Miller et al. 2004), and genes encoding proteins identified in mature *D. melanogaster* sperm are significantly co-localised (Dorus et al. 2006; Wasbrough et al. 2010). It should be noted that the model we present here was used to successfully identify clusters of genes encoding *D. melanogaster* sperm proteins (see Appendix IV: Wasbrough et al. 2010). In addition, large multi-gene domains of testis-specific genes have been observed in both Drosophila (Boutanaev et al. 2002) and *Mus. musculus* (mouse) (Divina et al. 2005; Li, Lee, and Zhang 2005). Finally, in Drosophila these testis neighbourhoods have been associated with B type lamin that tethers these domains to the nuclear envelope, a region correlated with repression of gene expression, in somatic but not testis cells (Shevelyov et al. 2009).

The eukaryotic genome is non-randomly organised and a variety of methods have been developed to characterize co-localised gene clusters and demonstrate their deviation from random patterns. However, many of these approaches make biologically unrealistic assumptions about neighbourhood characteristics, including pre-determining gene

neighbourhood size, or do not take directional biases in chromosome gene content into account. We aimed to construct a flexible algorithm for the identification of regions of co-localised genes that eliminates assumptions about neighbourhood size. To assess the utility of our model we investigated the co-localisation of testis-expressed genes in the *D. melanogaster* genome to demonstrate that definition of testis genes, minimum density of testis genes in neighbourhoods and directionality of neighbourhood search (5' to 3' vs. 3' to 5') affect the number, size and significance of neighbourhoods observed, and that this can effect observations of differences in both gene content and extent of gene co-localisation between the X chromosome and the autosomes. *D. melanogaster* was used for several reasons, primarily as it has a well annotated, highly studied genome, with high quality expression data available for 16 adult tissues and the whole fly (Chintapalli, Wang, and Dow 2007) and during spermatogenesis (Vibranovski et al. 2009a). Testis-expressed genes were targeted due to the availability of previous studies with which to compare our model, but also due to reported differences between the testis gene content of the autosomes and X chromosome. Significant under-representation of male-biased genes on the X chromosome compared the autosomal gene content have been reported, although the significance/extent of this difference depends on the definition of male biased (sex-biased, testis specificity) (Parisi et al. 2003; Sturgill et al. 2007) or gene age (Zhang et al. 2010). Clarifying the difference in gene content between the X chromosome and the autosomes, may also help determine the presence, or extent, of male germline meiotic sex chromosome inactivation, a controversial phenomenon in Drosophila (Hense, Baines, and Parsch 2007; Vibranovski et al. 2009a; Meiklejohn et al. 2011; Mikhaylova and Nurminsky 2011), and elucidate the drivers of gene duplication "off" the X chromosome (Betrán, Thornton, and Long 2002; Vibranovski, Zhang, and Long 2009b).

4.2 Materials and methods

4.2.1 NIM: Neighbourhood Identification Model

We developed an algorithm to identify co-localised genes (which we term neighbourhoods) that share a specific characteristic (*C*). Neighbourhoods were determined by two parameters: the minimum number (*N*) and the minimum density (*D*) of genes possessing *C*, within a genomic region. In addition, the outermost gene at each end of a neighbourhood (which we term a "boundary gene") must be a gene possessing *C*. As *D* can be less than 1, the algorithm is capable of identifying two types of neighbourhoods, which we term "tight" when D = 1, and "loose" when D <1.0. Genes are ordered along their respective chromosomes in the 5' to 3' orientation and the algorithm identifies the first and last gene on the chromosome that posses *C* (C_1 and *Cn*, respectively) (Figure 4.1). The algorithm then determines if the region between C_1 and *Cn* is meets the minimum requirements for a neighbourhood, as specified by *N* and *D*. If the region meets these requirements the genomic region is removed and the identified neighbourhood is stored in an array. However, if the region does not meet these parameters the algorithm identifies the next to last gene that posses *C* (*Cn-1*), and the process is repeated. This continues until either a neighbourhood is identified or $C_1 = Cn$. If $C_1 = Cn$, then C_1 is removed from the analysis, and the algorithm restarts from the next C_1 in the sequence. Once all neighbourhoods and non-neighbourhood regions have been removed the algorithm repeats in the opposite orientation (3' to 5'), to account for potential directional biases. Subsequently the model permits two options for reconciliation of neighbourhoods that do not contain the same genes when identified in opposite orientations: conservative and liberal (Figure 4.2). The conservative option returns neighbourhoods in which only genes that were found in the overlapping regions of the two sets of neighbourhoods, while the liberal option returns neighbourhoods that contain all genes identified as within neighbourhoods regardless of the orientation in which they were identified. The code for this algorithm is provided in accompanying CD.

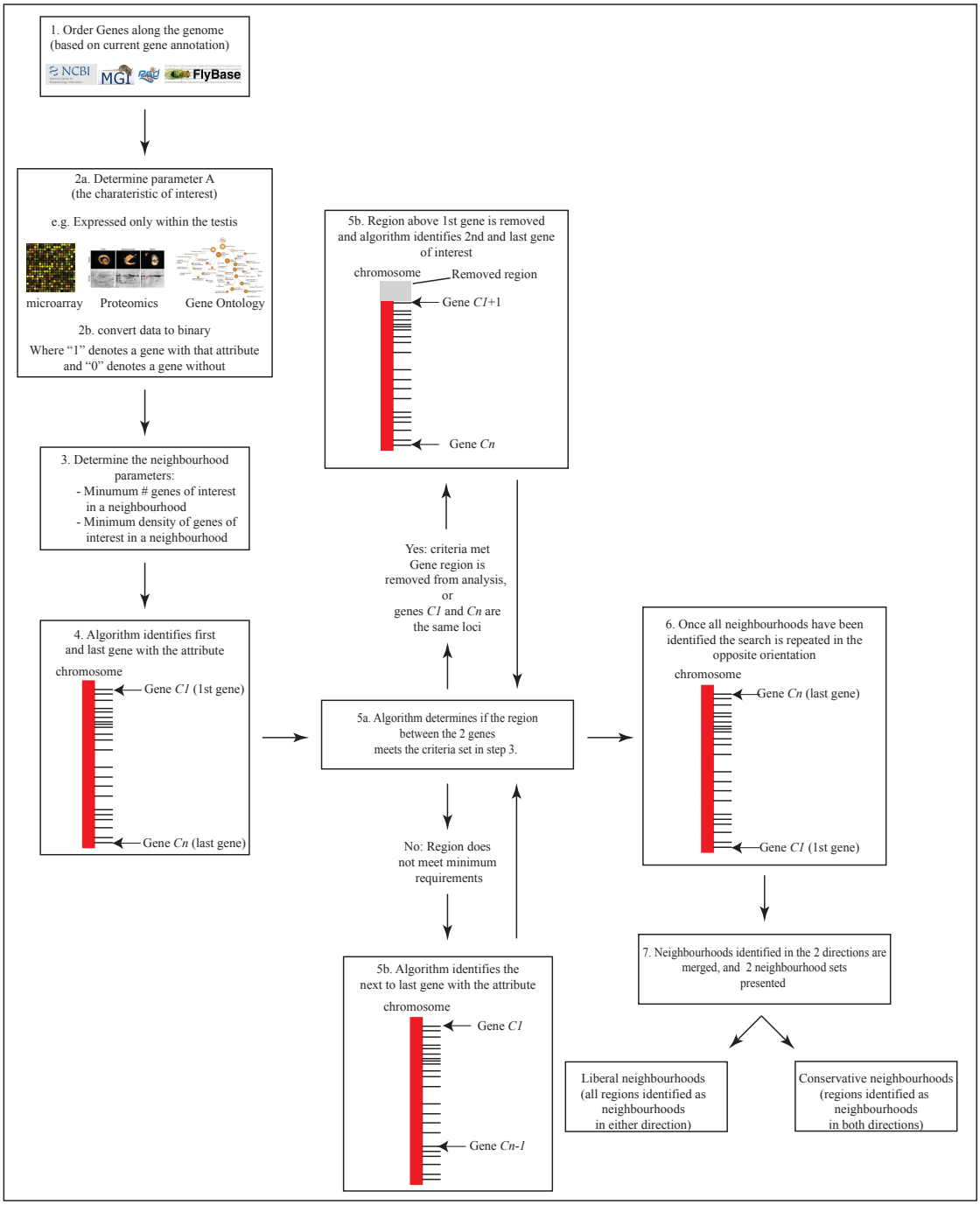

Figure 4.1 Diagram of neighbourhood identification algorithm.

Flow chart depicts how the neighbourhood identification model processes data.

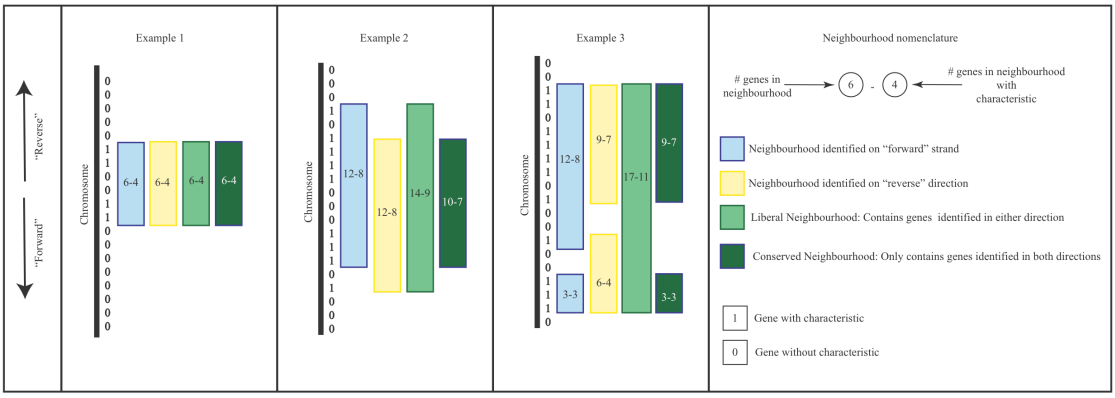

Figure 4.2 Resolution of neighbourhoods due to differences in directionality.

Genes possessing a characteristic of interest ("1") and genes without this characteristic ("0") have been ordered along the chromosome (black) and those identified as located within neighbourhoods have been highlighted by adjacent coloured blocks depending on the direction of the search: forward (blue) or reverse (yellow) strand direction. Neighbourhoods identified using the liberal (light green) and conservative (dark green) amalgamation methods are also provided. Liberal neighbourhoods contain all genes identified as residing in neighbourhoods in either orientation, while conservative neighbourhoods contain only genes identified as located in neighbourhoods in both orientations.

4.2.2 Model parameters

In our analysis of the co-localisation of testis-expressed genes in the *D. melanogaster* neighbourhood we consistently kept *N* as 3 genes in all analysis, while *D* was set as either 1 (tight neighbourhoods) or ≥ 0.66 (loose neighbourhoods). For loose neighbourhoods we compared the results of the different reconciliation methods (liberal and conservative). Finally, we explored the effect of different definitions of testis genes (*C*). We defined testis genes based on their (1) detectable expression in the testes and (2) enrichment of expression in testes relative to the remainder of the fly. Presence in a tissue was based upon detectable expression in at least 3 out of 4 experimental arrays in FlyAtlas (www.flyatlas.org). Utilising this criteria two testis gene sets were determined, the first contained all genes present in the testis (TP) and the second contained genes that were specifically present in the testis (TPS). Enrichment in a tissue was based on the statistical comparison of each tissue to the remainder of the fly that FlyAtlas provides. Genes could be classified as enriched in a tissue compared to the remainder of the fly (termed "Up" in FlyAtlas), no difference in expression between the tissue and the remainder of the fly (termed "None" in FlyAtlas) or decreased expression in the tissue compared to the remainder of the fly (termed "Down" in FlyAtlas). Based on this criteria we defined a further two sets of testis genes, the first contained all genes enriched in the testis (TE) and the second

contained genes that were only enriched in the testis (TES). Finally, we defined three additional sets of testis genes based on the ratio of their expression in the testis relative to the remainder of the fly. These sets contained genes that have 2-fold (T2), 5-fold (T5) or 10-fold (T10) higher expression in the testis compared to their expression in the fly.

4.2.3 Expression data curation

Microarray probe annotation (Drosophila_2.na32.annot.csv) was downloaded from Affymetrix, via the FlyAtlas website (19/3/2012). Probes for which no alignment location was provided (n = 188), probes that were aligned to multiple genome locations (n = 416), probes with no FlyBase gene symbol, FlyBase annotation symbol, FlyBase identifier or RefSeq identifier (n = 3,393), probes assigned to heterochromatic regions (n = 212), chromosome 4 (n = 103) or chromosome "U" (n = 55), and probes for which the assigned gene has been "withdrawn" from the FlyBase database (www.flybase.org) were removed from the dataset. Positions of all annotated genes within the *D. melanogaster* genome were obtained from FlyBase (19/3/2012; ftp file: gene_map_table_fb_2012_02.tsv). Remaining probes were checked to ensure the given probe alignment coordinates overlapped with part, or all, of the allocated gene. Probes for which the alignment coordinates did not overlap the assigned gene (n = 29) were removed from the dataset. In addition 10 probe sets were assigned to multiple genes that shared the same start and stop coordinates. These 10 probes were retained in the dataset. A further 16 probe sets were assigned to multiple genes that either partly overlapped or were found in tandem. These 16 probe sets were examined to discover how the probe overlapped these genes. Probes that were found to overlap both genes were removed from the dataset (n = 10). The resulting dataset consisted of 14,251 probe sets, which corresponded to 12,635 genes.

Microarray expression data was obtained as a tab-delimited file from FlyAtlas (3/4/2012). The probe annotation data and the expression data were matched based on the Affymetrix probe set identifier. Genes were ordered along their respective chromosome arms (2L, 2R, 3L, 3R and X) by start position. Genes that shared the same start co-ordinates were coalesced into one "genomic unit" and for all analysis treated as a single gene. The resulting dataset consisted of 13,547 current, protein coding, genomic units. Average gene expression signal was obtained for all genes (or genomic units) that were assigned more than one probe set. In addition to the expression signal, FlyAtlas also provides two additional measures: presence and enrichment. For each gene, or genomic unit, with multiple probe sets assigned we also obtained the average number of arrays with detectable expression (the measure of presence) and scored genes as enriched in a particular tissue if all probe sets had been classified as "up" in the tissue compared to the remainder of the fly. For genes with no probe/expression information assigned (n = 988), expression in each tissue, including the fly, was reported as 0, as was the number of arrays with

detectable presence, and for each tissues' enrichment was recorded as "none". The final datasets contained 8,419 TP genes, 347 TPS genes, 3,045 TE genes, 1,601 TES genes, 2,642 T2, 1,745 T5 and 1,156 T10 genes.

4.2.4 Statistical analysis

To assess the significance of co-localisation of testes genes we utilised non-parametric Monte Carlo simulations. Significance was based upon the number of simulations (n = 10,000), in which gene order was randomly assigned, that generated the same, or higher, number of neighbourhoods than the observed data, when the same combinations of parameters were applied. Directionality was not implemented as random data was assumed not to have a bias on the orientation in which the search was performed. Therefore for loose neighbourhoods, which had two separate amalgamation methods (liberal and conservative) to resolve differences due to model directions, the smallest number of neighbourhoods was used to determine a single p-value for all loose neighbourhoods at that density and testis gene set. Differences in proportion of testis genes between the X chromosome and autosomes and between the different sets of testis genes, and differences in the proportion of testis genes within neighbourhoods between the X chromosome and the autosomes and between the different sets of testis genes were assessed using χ^2 tests with Yates correction, and Bonferroni correction for multiple testing where necessary. Kolmogorov-smirnov tests were employed to assess differences in the range of neighbourhood sizes between the X chromosome and autosomes, and between the different sets of testis genes.

4.3 Results

4.3.1 Significant co-localisation of testis genes in the D. melanogaster genome

We observed a significantly higher number of tight testis neighbourhood, on both the autosomes and X chromosome, for all testis gene sets, than expected by random rearrangement of the genome (Table 4.1). Similarly, we observed that the number of loose testis neighbourhoods is significantly higher than expected, with one exception: loose testis neighbourhoods do not occur significantly more than expected in the TE testis gene dataset (Table 4.2). It should be noted that the resolution of orientation has produced a small number of conserved neighbourhoods consisting of only 2 adjacent genes, which were removed from this dataset, and several liberal neighbourhoods that have a lower density of testis genes than permitted by *D*, which were retained in the dataset. Therefore the significance of co-localisation was obtained using the lower number of loose neighbourhoods. In both tight and loose neighbourhood models, the number of neighbourhoods differs between the testis gene sets, although this is likely an artefact of the differing numbers of genes within each dataset. Similarly, while the numbers of loose and tight neighbourhoods are more numerous on the autosomes compared to the X chromosome, this is largely attributable to difference in the number of X-linked and autosomal genes, as both have similar proportions of testis genes residing on them.

Predictably, due to the lower stringency of the neighbourhood parameters, loose testis neighbourhoods are more numerous than tight testis neighbourhoods. The number of neighbourhoods identified with $D = 0.66$ has increased by ~30-50% depending on the testis dataset compared to the number of neighbourhoods identified with $D = 1$ (TE: 52.8%, TES: 32.6%, T2: 52.3%, T5: 42.2%, T10: 36.4%; increases based on number of conservative loose testis neighbourhoods). Although, there is significantly more co-localisation of testis genes than expected, the majority of testis genes are not located within tight (>75%) or loose (>50%) testis neighbourhoods. As expected loose testis neighbourhoods contain a higher proportion of testis genes than tight testis neighbourhoods. Compared to tight testis neighbourhoods, loose neighbourhoods contain ~10-20% more testis genes (TE: 18.7%, TES: 12.7%, T2: 18.3%, T5: 13.0%, T10: 9.4%; increases are based on the difference in proportion of testis genes in tight neighbourhoods and loose conservative neighbourhoods).

Table 4.1 Genomic enrichment of testis expression neighbourhoods (D = 1.0)

Criteria*	Autosomes			X Chromosome		
	Observed	Expected†	*p*	Observed	Expected†	*p*
TE	164	111.4	< 0.0001	16	9.9	0.0323
TES	77	17.5	< 0.0001	9	2.3	0.0005
T2	132	74.3	< 0.0001	19	7.6	0.0001
T5	70	21.8	< 0.0001	13	3.3	< 0.0001
T10	40	6.7	< 0.0001	4	1.0	0.0151

* TE- Testis expressed; TES- Specifically Testis Expressed; T2/T5/T10 - 2-, 5-, 10- fold higher expression in the testis compared to the remainder of the fly, respectively

† The average number of neighbourhoods stochastically generated

Table 4.2 Genomic enrichment of testis expression neighbourhoods (*D* = 0.66)

Criteria*	Autosomes			X Chromosome		
	Observed§	Expected†	*p*	Observed§	Expected†	*p*
TE	249/251	232.1	0.0519	26/26	23.2	0.2609
TES	100/102	45.1	< 0.0001	14/14	6.0	0.0013
T2	204/206	165.7	< 0.0001	26/25	18.3	0.0379
T5	100/102	55.3	< 0.0001	18/17	8.5	0.0018
T10	50/51	18.1	< 0.0001	10/10	2.7	0.0002

§Number of Conservative loose neighbourhoods/ number of Liberal loose neighbourhoods

* TE- Testis expressed; TES- Specifically Testis Expressed; T2/T5/T10 - 2-, 5-, 10- fold higher expression in the testis compared to the remainder of the fly, respectively

† The average number of neighbourhoods stochastically generated

We observed that tight testis neighbourhoods contained 3 to 10 testis genes, while loose testis neighbourhoods could be much larger and ranged in size from 3 to >30 genes. For tight testis neighbourhoods we observed no significant differences between the distributions of neighbourhood sizes between testis gene datasets. In all testis gene datasets the median, and inter quartile ranges of tight neighbourhood sizes were the same (Q_1 = 3 genes, Q_2 = 3 genes, Q_3 = 4 genes) with the exception of tight testis neighbourhoods on the X chromosome using the T10 testis gene dataset (Q_1 = 3 genes, Q_2 = 4 genes, Q_3 = 5.5 genes). Similarly, we observed no significant differences in the distribution of loose testis neighbourhood sizes between different

testis gene datasets (Figure 4.3). Median loose neighbourhood sizes were larger than tight testis neighbourhoods, for both liberal and conservative loose neighbourhoods. Similarly median neighbourhood size was marginally larger for liberal loose neighbourhoods compared to conservative loose neighbourhoods.

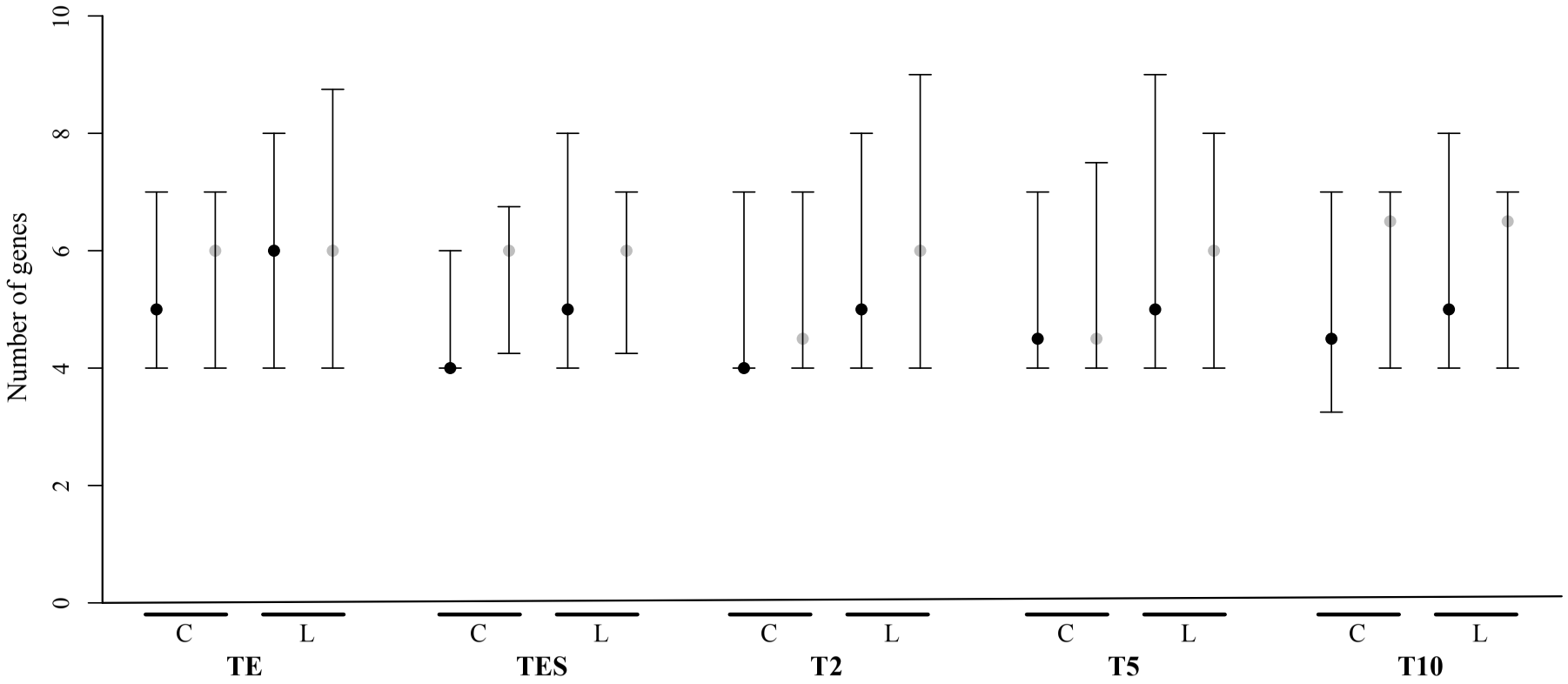

Figure 4.3 Median size of loose testis genes neighbourhoods.

Median number of testis genes (TE - testis enriched; TES - specifically testis enriched; T2, T5, T10 - 2-, 5-, 10-fold higher expression in the testis compared to the remainder of the fly, respectively) in conservative (C) and liberal (L) loose neighbourhoods on the autosomes (black) and X chromosome (grey) ± interquartile range.

4.3.2 Large testis neighbourhood located on chromosome 2

Previous studies have identified large regions of co-localised testis genes on chromosome 2 (Boutanaev et al. 2002). Consistently, in all testis gene datasets we observe a large loose testis neighbourhood (~270kb) on chromosome arm 2R (Figure 4.4). This region contains 39 genes, of which 28 are in the TE dataset, 26 are in the TES dataset, and 27, 26 and 22 are in the T2, T5 and T10 datasets, respectively. This region includes 2 genes (*ord* and *CG30410*) previously identified in the *D. melanogaster* sperm proteome (Dorus et al. 2006; Wasbrough et al. 2010). In addition, there are 3 other genes (*CG34393, CG3085* and *Fib*) that also encode novel sperm components within 10 genes of this region. Only *ord* has been implicated as having a sterility phenotype (www.flybase.org). Many of the genes in this region have no gene ontology categories assigned to them in AmiGO (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi). However, three genes appear to be involved in proteolysis. *CG3502* is implicated in aminopeptidase (GO:0004177) and metallopeptidase (GO:0008237) activity, while *Prosbeta5R* and *Yip3* are both implicated in threonine-type endopeptidase activity (GO:0004298) and proteolysis involved in cellular protein catabolic process. (GO:0004298).

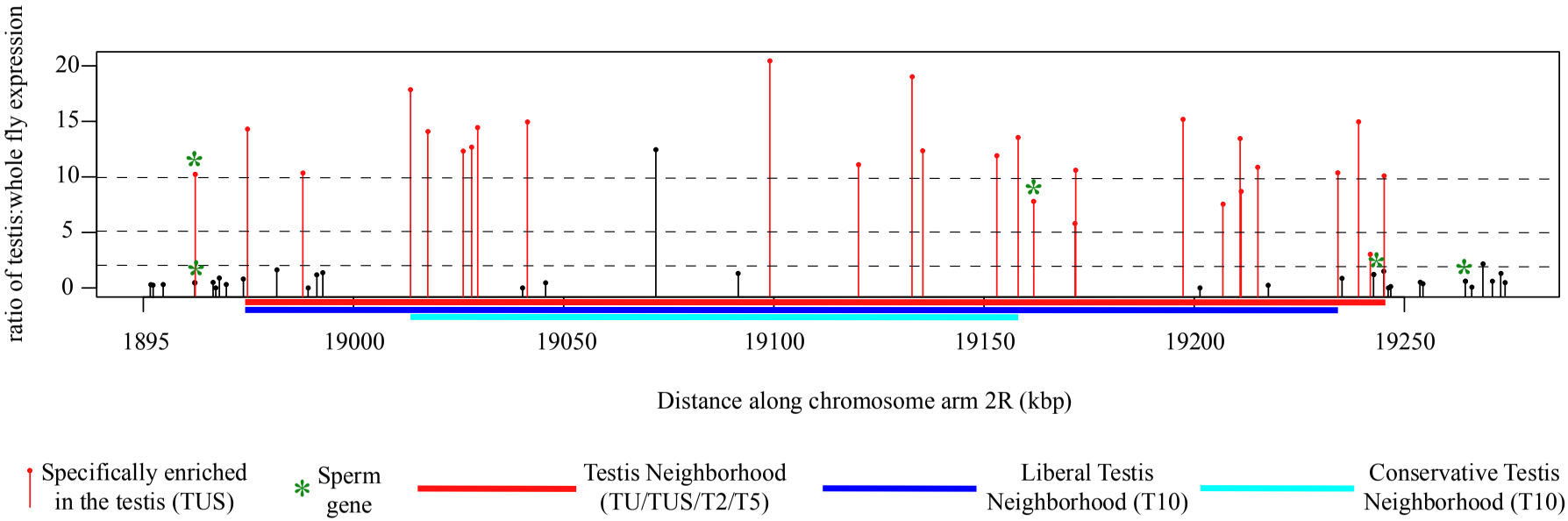

Figure 4.4 Large testis neighbourhood located on chromosome arm 2R.

Genes are located along chromosome 2R by their start position. Each genes relative expression in the testis (based upon the ratio of expression in the testis compared to the whole fly, provided by FlyAtlas [www.flyatlas.org]) is provided. Genes specifically enriched within the testis (TES) are highlighted in red. Genes empirically identified as encoding a novel protein component of mature spermatozoa by mass spectrometry (Dorus et al. 2006; Wasbrough et al. 2010) were indicated by a green asterisk. The span of the testis neighbourhood based upon (a) loose co-localization of gene testis enriched (TE), testis specifically enriched (TES), and 2-fold and 5-fold higher expression in the testis compared to the whole fly (red), (b) liberal co-localization (dark blue) and (c) conservative co-localization (light blue) of genes with 10-fold higher testis expression compared to the whole fly is highlighted. Only loose co-localization of TE, TES, T2 and T5 is provided as the same region was identified using all four categories and in both directions.

4.3.3 Comparison between the X chromosome and the autosomes

Due to studies demonstrating the difference in gene content, particularly regarding male-biased genes, between the X chromosome and autosomes, we compared the testis gene content and extent of co-localisation on the X chromosome and autosomes. We observed no significant difference in the proportion of genes without data on the X chromosome (6.99%) and on the autosomes (8.03%) ($\chi^2 = 2.075$; $p = 0.1497$). In general we observed no significant difference between the proportion of X chromosome and autosome genes in the testis gene dataset, with two exceptions. We observed a significantly higher proportion of TE genes resided on the autosome, compared to the X chromosome ($\chi^2 = 32.119$; $p < 0.0001$). This was also observed in the T2 dataset ($\chi^2 = 17.972$; $p < 0.0001$) but not in any other testis gene set (Figure 4.5).

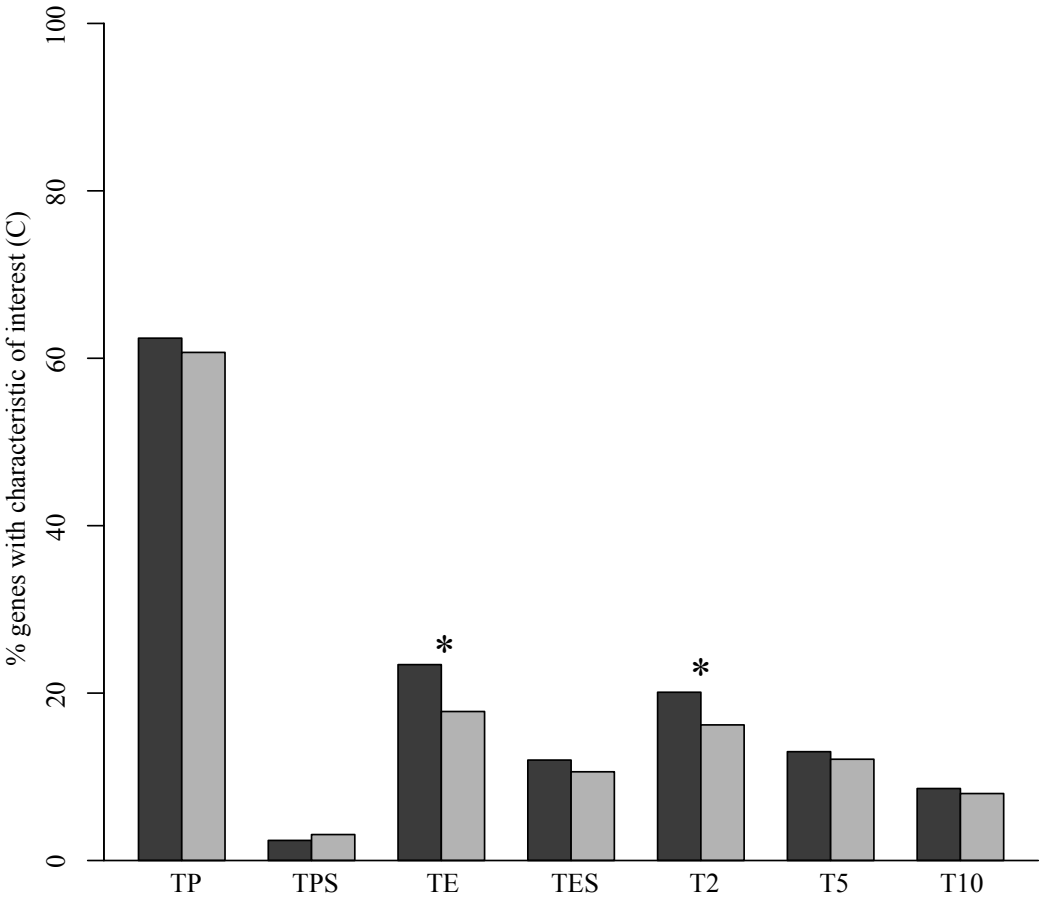

Figure 4.5 Proportion of genome classified as testis genes.

Percentage of genes on autosomes (dark grey) and the X chromosome (light grey) that have been defined as TP (present in the testis), TPS (only present in the testis) TE (testis enriched), TES (specifically testis enriched), T2, T5 or T10 (2, 5, or 10 fold higher expression in the testis compared to the remainder of the fly) Asterisk (*) denotes significant difference in the proportion of genes identified as testis genes located on the autosomes and chromosome X.

We observed no tendency for different proportions of testis genes to reside in tight (Figure 4.6) or loose (Figure 4.7) testis neighbourhoods between the X chromosome and autosomes. However, we observed a significantly smaller proportion of X-linked, compared to autosomally-linked, testis genes residing in tight testis neighbourhoods using the TE definition ($p = 0.0047$). Similarly, we observed a significant difference in the proportion of testis genes residing within loose neighbourhoods between the X chromosome and the autosomes for TE (conservative: $\chi^2 = 14.370$, $p = 0.0002$; liberal: $\chi^2 = 17.798$, $p < 0.0001$); however, unlike in tight testis neighbourhoods we also observed this for T2 (conservative: $\chi^2 = 5.307$, $p = 0.0212$; liberal: $\chi^2 = 4.428$, $p = 0.0354$). Finally, we observed no significant difference in the distribution of neighbourhood sizes between neighbourhoods on the autosomes and the X chromosomes, for both tight and loose testis neighbourhoods.

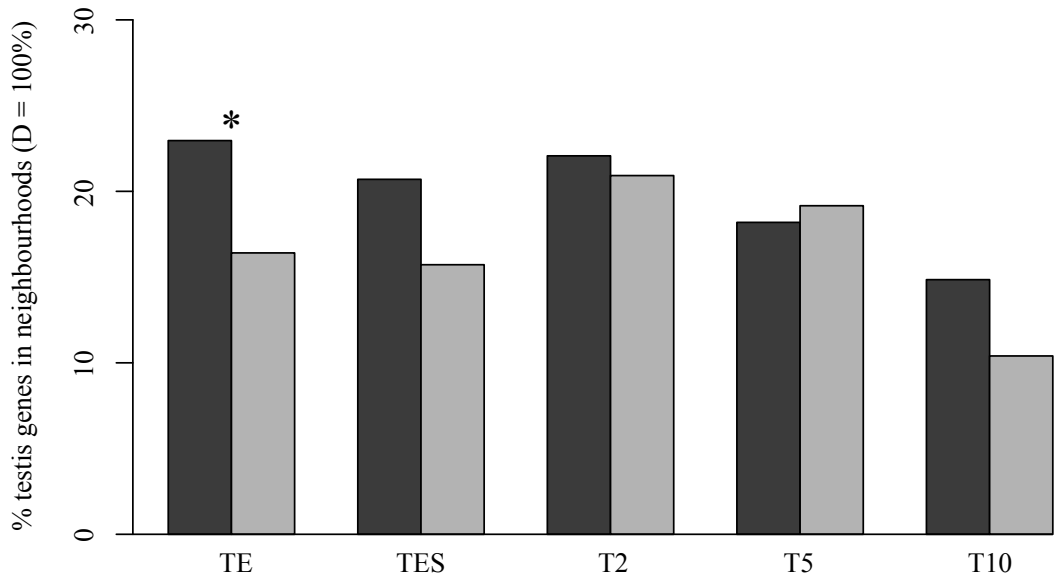

Figure 4.6 Proportion of testis genes co-localised into tight neighbourhoods.

The percentage of testis genes (TE - testis enriched; TES - specifically testis enriched; T2, T5, T10 - 2-, 5- or 10-fold higher expression in the testis compared to the remainder of the fly) located within tight neighbourhoods on the autosomes (dark grey) and X chromosome (light grey). Asterisk (*) denotes significant difference in the proportion of genes identified as testis genes located on the autosomes and chromosome X.

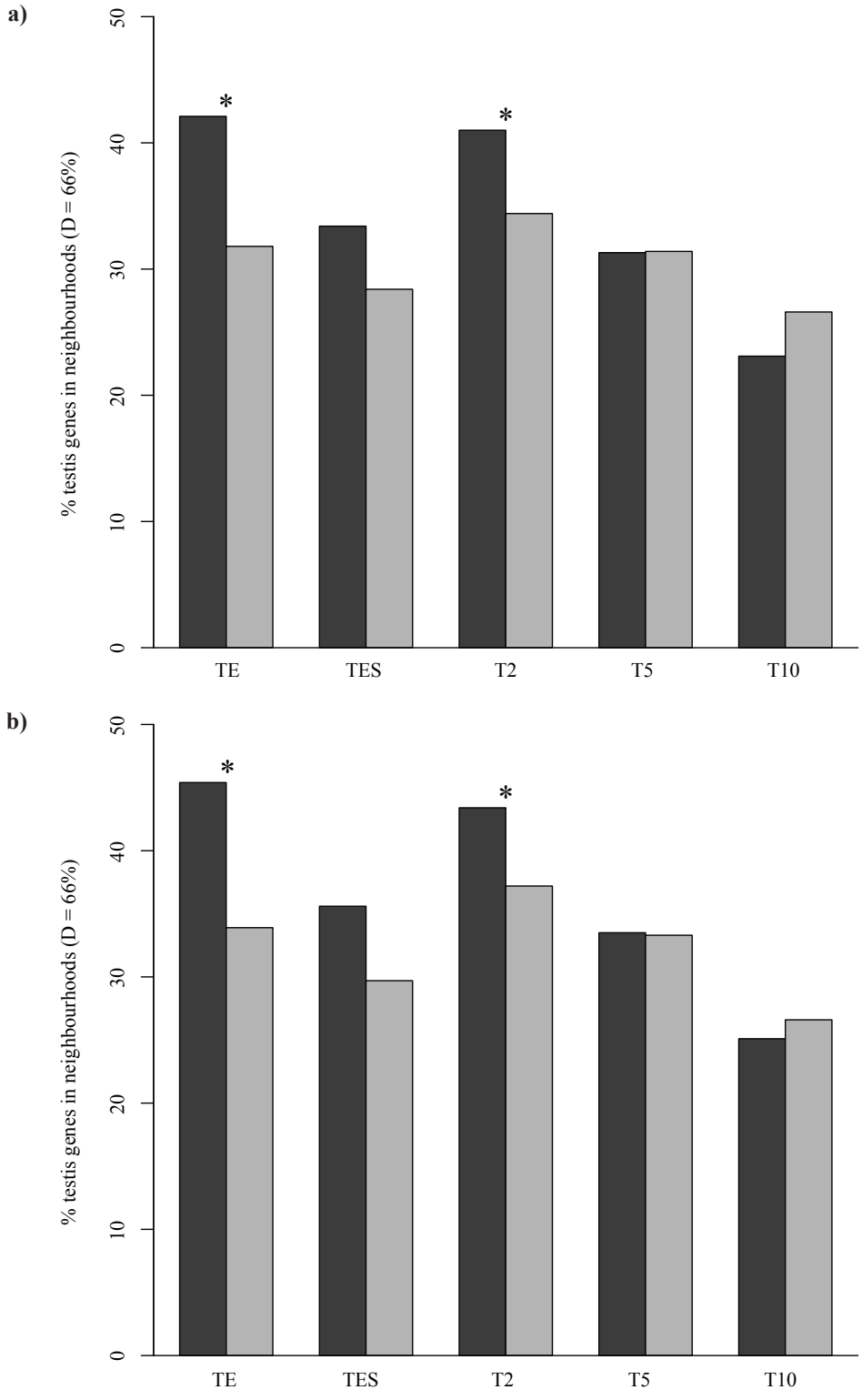

Figure 4.7 Proportion of testis genes co-localised into loose neighbourhoods.

The percentage of testis genes (TE - testis enriched; TES - specifically testis enriched; T2, T5, T10 - 2-, 5-, 10- fold higher expression in the testis compared to the remainder of the fly, respectively) located within (a) conservative and (b) liberal loose neighbourhoods, on the autosomes (dark grey) and X chromosome (light grey). Asterisk (*) denotes significant difference in the proportion of genes identified as testis genes located on the autosomes and chromosome X.

4.4 Discussion

The eukaryotic genome is not randomly organised with regards to gene location (Hurst, Pál, and Lercher 2004). Genes have been observed to be spatially organised based on their expression. Several methods have been devised to assess the significance of this spatial colocalization, however many of these methods have biologically unrealistic parameters. In this chapter we present a new model for the identification of gene neighbourhoods, which is flexible, easy to implement and requires few prior assumptions about the nature of gene neighbourhoods. Utilising this model we re-analysed the co-localisation of testis expressed genes in the *D. melanogaster* genome, which confirmed the presence of significant multi-gene neighbourhoods similar to those previously documented (Boutanaev et al. 2002). However our analysis also highlighted potential confounding factors such as the definition of testis expressed genes and setting of the minimum density of testis genes within neighbourhoods, which need to be considered in any study of co-localisation of genes as they can affect observations regarding both the overall significance of co-localisation as well as distinctions between the X chromosome and the autosomes.

4.4.1 Evaluation of NIM (Neighbourhood Identification model) performance

Similar to the previous analysis of the spatial distribution of testis genes in *D. melanogaster* (Boutanaev et al. 2002) we observed significant co-localisation of testis expressed genes in the *D. melanogaster* genome. It is difficult to directly compare our results due to differences in genome annotation. However, despite this we observe two general consistencies with studies of testis gene neighbourhoods in both *D. melanogaster* (Boutanaev et al. 2002) and mouse (Divina et al. 2005; Li, Lee, and Zhang 2005): (1) all studies observed greater co-localisation of testis genes than expected by stochastic distribution of gene position and (2) all studies observed that the majority of testis genes were not located in testis neighbourhoods. We observed that 20-40% of testis genes resided within testis neighbourhoods, which is similar to Boutanaev *et al*. (2002) in which ~1/3 of testis genes were identified in neighbourhoods. Many of the differences between our study and Boutanaev *et al* (2002) are potentially due to differences in how we defined a "testis gene", improvements in gene annotation and expression datasets and that we allowed non-testis genes in (some) of our neighbourhoods. Despite these differences, in general our model has reported results consistent with previous studies.

Our newly developed model for the identification of gene neighbourhoods is simple, requires no complex transformations of the data or intensive computational resources, and can be used on a variety of data types. More importantly our model requires few assumptions regarding the data including: the symmetry of genes along chromosomes or within neighbourhoods, the proximity of neighbourhoods to each other, the distribution of neighbourhoods along chromosomes or

neighbourhood size. Further, our proviso that the neighbourhood boundary genes must possess *C* means that our neighbourhoods have a defined, but not specified *a priori*, beginning and end point. This is an improvement over previous methods such as the sliding window approach in which the neighbourhood may extend to the end of the window but the boundary gene, and several of its neighbours, may not possess *C*. Finally, to our knowledge, no previous model has taken the directionality in which the search is conducted into account. Therefore by default they have assumed that both neighbourhoods and the chromosome have no bias in density or distribution of genes possessing *C*. We demonstrate that depending on whether the search for neighbourhoods is conducted in a 5' to 3' or a 3' to 5' orientation can alter the genes that are deemed to reside in neighbourhoods. Furthermore we provide a conservative and non-conservative (liberal) method of reconciling these disparities.

However, the model has an intrinsic bias to identify larger, less dense, neighbourhoods over smaller, highly dense, neighbourhoods; although this can be largely resolved by the employment of the conservative amalgamation method for the resolution of orientation differences. However, each of the amalgamation methods has a defined disadvantage. Due to our flexibility in allowing genes possessing *C* to be non-symmetrically dispersed within our loose neighbourhoods we observe that some neighbourhoods may be artificially extended by individual outlying genes. For example, there may be a core cluster of the majority of the *C* genes at the 5' end of the neighbourhood, while a single *C* gene may expand the neighbourhood towards the 3' end, increasing overall neighbourhood size while decreasing neighbourhood *C* density. This is more pronounced when the liberal reconciliation method is applied and can lead to neighbourhoods with a minimum density of *C* below the specified *D*. Conversely, the application of the conservative reconciliation method results in smaller, denser neighbourhoods, but ignores outlying *C* genes. It is also possible that by removing these outliers using the conservative reconciliation method we may be artificially removing regions that are still controlled by that chromatin domain, these genes may not be near the "core" but are in close enough proximity to still be regulated by the chromatin domain. An improvement to the model to address this may be the addition of a third parameter: distance. This may take the form of the maximum number of genes that can be found between the boundary genes and the nearest gene possessing *C* within the neighbourhood (Weber and Hurst 2011) or a maximum intervening genomic distance between these two genes (Li, Lee, and Zhang 2005). However, knowing the stringency with which to apply a parameter of distance requires further investigation by ChIP in order to determine the average size of open and closed chromatin regions. Finally this model, and all other methods, cannot account for 3-dimensional movements of chromosomes within the nucleus or the formation of multi-chromosome spanning neighbourhoods created by changes in the proximity of genomic regions.

4.4.2 Characteristics of testis neighbourhoods

As we have stated, we observed significantly greater co-localisation of testis genes than expected by random simulation of the genome, with the exception of loose neighbourhoods based on the TE definition of a "testis gene". Despite the extensive overlap in genes identified by each definition of "testis genes", many of the differences in the number of neighbourhoods identified and the proportion of testis genes within neighbourhoods can be attributed to differences between these "testis gene" definitions. However, while the range in size of neighbourhoods' ranges from 3 to over 30 genes the distribution of neighbourhood sizes between "testis gene" neighbourhoods does not significantly alter. In addition to differences based on different definitions of a testis gene, we also observed differences based on variations in the specified *D*. Predictably as *D* decreased the number and size of neighbourhoods, and the proportion of testis genes residing in testis neighbourhoods, increased. However, it is noteworthy that no scenario resulted in >50% of testis genes co-localising into neighbourhoods, and this is consistent with other studies. This could be due to several, not mutually exclusive, possibilities, including: that it is not possible to observed genome organisation of genes in a simple linear model due to the 3-dimensional complexity of nuclear architecture (Lanctôt et al. 2007; Splinter and De Laat 2011). A linear model cannot observe 3-dinemsional neighbourhoods created by the bringing together of different chromosome sections within the nucleus at specific times. It is also possible that testis expressed genes may be co-localised with non-testis expressed genes because both have similar spatio-temporal expression profiles in other tissues or simply because these non-testis genes may not be defined as "testis genes" as they have low levels of testis expression or similar levels of expression in the testis and many other tissues. Finally, it may be because we have only identifies the "core" of the neighbourhood, and the effect of the chromatin domain in terms of both regulation and selection for similar expressed genes may extend beyond this region. In support of the later proposal, we observed a large testis neighbourhood on chromosome 2. Within this neighbourhood are two genes that encode mature sperm proteins (*ord* and *CG30410*) as well as several genes that function in proteolysis, a process that may be important in reproduction (see chapter 2). However, nearby this neighbourhood are several genes with testis expression, and three genes that encode sperm proteins (*CG34393, CG3085* and *Fib*), suggesting that higher-order regulation and selection extend beyond the region identified as a testis neighbourhood.

However, if we have identified only the "core" of each chromatin domain this should include the transcription initiation factors used to open chromatin and begin transcription. As such disruptions in this regions should effect gene expression within this region. Testis gene neighbourhoods in Drosophila have been disrupted using chromosomal inversions, but this did

not affect gene expression within the neighbourhood, even when the inversions lead to the spatial separation of neighbourhoods in the nucleus (Meadows et al. 2010). This suggests that chromatin domains are malleable, and can survive even large-scale genome rearrangements, which may be of interest when studying the evolutionary conservation of these neighbourhoods. Meadows *et al.* (2010) separated each testis neighbourhood into two halves. It is possible that these halves were still large enough to contain initiation factors and therefore function independently. The neighbourhoods observed here consistently, regardless of definition of *C*, range from 3 to over 30 genes. It is possible that there is a minimum number of genes required for a neighbourhood to act independently. However it may be that this minimum neighbourhood size requires 3-dimensional modelling and ChiP technology to follow the position of multiple testis neighbourhoods and assess whether they converge to form meta-neighbourhoods, and whether the neighbourhoods forming these meta-neighbourhoods are always consistent.

4.4.3 Differences between the X chromosome and autosomes differ depending on definitions

In contrast both to theory and previous observations that the X chromosome should differ from the autosomes in terms of male-biased gene content (Gurbich and Bachtrog 2008; Meisel, Malone, and Clark 2012), we observe few differences in (1) the proportion of X-linked and autosomal genes that are testis-expressed, (2) the proportion of testis genes residing in X-linked or autosomal testis neighbourhoods, (3) the significance of co-localisation of testis genes, or (4) the distribution of neighbourhood sizes. While fewer neighbourhoods and testis genes are located on the X chromosome compared to the autosomes, this is primarily associated with the overall differences in size of the X chromosome compared the collective autosomes. The few difference we observe between the X chromosome and the autosomes occur within the TE (testis enriched) and T2 (2-fold higher expression in the testis compared to the remainder of the whole fly) gene sets. Differences may occur in these sets because there are more genes within these gene sets, giving us a larger sample size and therefore resolving differences. Alternatively, as differences between the X chromosome and autosomes appear to disappear as we increased either specificity to the testis (TES) or level of expression within the testis (T5, T10), it is possible that differences occur at lower levels of testis expression, or specificity, but not at higher, because selection is operating on low level testis expressed genes. These genes (TE, T2) are not necessarily only expressed in the testis, and therefore there is a greater potential that these genes are sexually antagonistic, and therefore it is disadvantageous for these genes to be X-linked resulting in the significantly lower proportion of T2 and TE genes on the X chromosome. Alternatively, differences may not appear between the X chromosome and the autosomes at higher levels of testis expression, or specificity, as it does not matter if these genes are on the X chromosome as they will not be expressed in the soma and are therefore not

sexually antagonistic. Similarly because highly testis specific, or expressed, genes are not required in the soma there may be selection for them to co-localise into neighbourhoods, regardless of whether they are X or autosomally linked, so that they can be silenced in the soma but effectively transcribed in the testis.

4.4.4 Summary

The *D. melanogaster* genome is organised, with testis-expressed genes tending to co-localise into gene neighbourhoods more often than expected by stochastic organisation of the genome. Our new model demonstrates consistent results with previous studies. However, we suggest that this, and previous models, have only identified the "core" genomic regions that are controlled, and influenced, by chromatin regulations and that nearby regions are also under similar, albeit relaxed, control. Furthermore, our analysis suggests that future studies need to be mindful in how they define a tissue-expressed/specific gene. Although this definition appears to have no effect on the distribution or median neighbourhood sizes, it may have implications on the number of neighbourhoods identified, the significance of co-localisation and differences between the sex chromosomes and autosomes.

4.5 Acknowledgments

I would like to thank Marios Richards and Dr. Kirill Borziak for creating the programme with which to identify gene neighbourhoods, without which this work and that of the next chapter would not have been possible.

4.6 References

Batada, Nizar N, and Laurence D Hurst. 2007. “Evolution of Chromosome Organization Driven by Selection for Reduced Gene Expression Noise.” *Nature Genetics* 39 (8): 945–949.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. “Retroposed New Genes Out of the X in Drosophila.” *Genome Research* 12 (12): 1854–1859.

Boutanaev, Alexander M, Aila I Kalmykova, Yuri Y Shevelyov, and Nurminsky Dmitry I. 2002. “Large Clusters of Co-expressed Genes in the Drosophila Genome.” *Nature* 420 (6916): 666–669.

Cairns, Bradley R. 2009. “The Logic of Chromatin Architecture and Remodelling at Promoters.” *Nature* 461 (7261): 193–198.

Celniker, Susan E, Laura A L Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, et al. 2009. “Unlocking the Secrets of the Genome.” *Nature* 459 (7249): 927–930.

Chiaromonte, Francesca, Webb Miller, and Eric E Bouhassira. 2003. “Gene Length and Proximity to Neighbors Affect Genome-Wide Expression Levels.” *Genome Research* 13 (12): 2602–2608.

Chintapalli, Venkateswara R, Jing Wang, and Julian A T Dow. 2007. “Using FlyAtlas to Identify Better *Drosophila melanogaster* Models of Human Disease.” *Nature Genetics* 39 (6): 715–720.

Divina, Petr, Cestmír Vlcek, Petr Strnad, Václav Paces, and Jirí Forejt. 2005. “Global Transcriptome Analysis of the C57BL/6J Mouse Testis by SAGE: Evidence for Nonrandom Gene Order.” *BMC Genomics* 6: 29.

Dorus, Steve, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L Karr. 2006. “Genomic and Functional Evolution of the *Drosophila melanogaster* Sperm Proteome.” *Nature Genetics* 38 (12): 1440–1445.

Garcia-Fernàndez, Jordi. 2005. “The Genesis and Evolution of Homeobox Gene Clusters.” *Nature Reviews. Genetics* 6 (12): 881–892.

Gurbich, Tatiana A, and Doris Bachtrog. 2008. “Gene Content Evolution on the X Chromosome.” *Current Opinion in Genetics & Development* 18 (6): 493–498.

Hense, Winfried, John F Baines, and John Parsch. 2007. “X Chromosome Inactivation During Drosophila Spermatogenesis.” *PLoS Biology* 5 (10): 2288–2295.

Hurst, Laurence D, Csaba Pál, and Martin J Lercher. 2004. “The Evolutionary Dynamics of Eukaryotic Gene Order.” *Nature Reviews. Genetics* 5 (4): 299–310.

International Human Genome Sequencing Consortium. 2004. “Finishing the Euchromatic Sequence of the Human Genome.” *Nature* 431 (7011): 931–945.

Kaiser, Vera B, and Hans Ellegren. 2006. “Nonrandom Distribution of Genes with Sex-biased Expression in the Chicken Genome.” *Evolution* 60 (9): 1945–1951.

Khil, Pavel P, Natalya A Smirnova, Peter J Romanienko, and R Daniel Camerini-Otero. 2004. “The Mouse X Chromosome Is Enriched for Sex-biased Genes Not Subject to Selection by Meiotic Sex Chromosome Inactivation.” *Nature Genetics* 36 (6): 642–646.

Koonin, Eugene V. 2009. “Evolution of Genome Architecture.” *The International Journal of Biochemistry & Cell Biology* 41 (2): 298–306.

Lanctôt, Christian, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. 2007. “Dynamic Genome Architecture in the Nuclear Space: Regulation of Gene Expression in Three Dimensions.” *Nature Reviews. Genetics* 8 (2): 104–115.

Lelli, Katherine M, Matthew Slattery, and Richard S Mann. 2012. “Disentangling the Many Layers of Eukaryotic Transcriptional Regulation.” *Annual Review of Genetics* 46: 43–68.

Lercher, Martin J, and Laurence D Hurst. 2006. “Co-expressed Yeast Genes Cluster over a Long Range but Are Not Regularly Spaced.” *Journal of Molecular Biology* 359 (3): 825–831.

Lercher, Martin J, Araxi O Urrutia, and Laurence D Hurst. 2002. “Clustering of Housekeeping Genes Provides a Unified Model of Gene Order in the Human Genome.” *Nature Genetics* 31 (2): 180–183.

Lercher, Martin J, Araxi O Urrutia, and Laurence D Hurst. 2003. “Evidence That the Human X Chromosome Is Enriched for Male-specific but Not Female-specific Genes.” *Molecular Biology and Evolution* 20 (7): 1113–1116.

Li, Quan, Bernett T K Lee, and Louxin Zhang. 2005. “Genome-scale Analysis of Positional Clustering of Mouse Testis-specific Genes.” *BMC Genomics* 6: 7.

Loppin, Benjamin, David Lepetit, Steve Dorus, Pierre Couble, and Timothy L Karr. 2005. “Origin and Neofunctionalization of a Drosophila Paternal Effect Gene Essential for Zygote Viability.” *Current Biology* 15 (2): 87–93.

Meadows, Lisa A, Yuk Sang Chan, John Roote, and Steven Russell. 2010. “Neighbourhood Continuity Is Not Required for Correct Testis Gene Expression in Drosophila.” *PLoS Biology* 8 (11): e1000552.

Meiklejohn, Colin D, Emily L Landeen, Jodi M Cook, Sarah B Kingan, and Daven C Presgraves. 2011. “Sex Chromosome-specific Regulation in the Drosophila Male

Germline but Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation." *PLoS Biology* 9 (8): e1001126.

Meisel, Richard P, John H Malone, and Andrew G Clark. 2012. "Disentangling the Relationship Between Sex-biased Gene Expression and X-linkage." *Genome Research* 22 (7): 1255–1265.

Michalak, Pawel. 2008. "Coexpression, Coregulation, and Cofunctionality of Neighboring Genes in Eukaryotic Genomes." *Genomics* 91 (3): 243–248.

Mikhaylova, Lyudmila M, and Dmitry I Nurminsky. 2011. "Lack of Global Meiotic Sex Chromosome Inactivation, and Paucity of Tissue-specific Gene Expression on the Drosophila X Chromosome." *BMC Biology* 9: 29.

Miller, Michael A, Asher D Cutter, Ikuko Yamamoto, Samuel Ward, and David Greenstein. 2004. "Clustered Organization of Reproductive Genes in the *C. elegans* Genome." *Current Biology* 14 (14): 1284–1290.

Mueller, Jacob L, Shantha K Mahadevaiah, Peter J Park, Peter E Warburton, David C Page, and James M A Turner. 2008. "The Mouse X Chromosome Is Enriched for Multicopy Testis Genes Showing Postmeiotic Expression." *Nature Genetics* 40 (6): 794–799.

Ng, Yen Kaow, Wei Wu, and Louxin Zhang. 2009. "Positive Correlation Between Gene Coexpression and Positional Clustering in the Zebrafish Genome." *BMC Genomics* 10: 42.

Oliver, Brian, and Tom Misteli. 2005. "A Non-random Walk Through the Genome." *Genome Biology* 6 (4): 214.

Parada, Luis A, Philip G McQueen, and Tom Misteli. 2004. "Tissue-specific Spatial Organization of Genomes." *Genome Biology* 5 (7): R44.

Parisi, Michael, Rachel Nuttall, Daniel Naiman, Gerard Bouffard, James Malley, Justen Andrews, Scott Eastman, and Brian Oliver. 2003. "Paucity of Genes on the Drosophila X Chromosome Showing Male-biased Expression." *Science* 299 (5607): 697–700.

Schneider, Robert, and Rudolf Grosschedl. 2007. "Dynamics and Interplay of Nuclear Architecture, Genome Organization, and Gene Expression." *Genes & Development* 21 (23): 3027–3043.

Shevelyov, Y Y, S A Lavrov, L M Mikhaylova, I D Nurminsky, R J Kulathinal, K S Egorova, Y M Rozovsky, and D I Nurminsky. 2009. "The B-type Lamin Is Required for Somatic Repression of Testis-specific Gene Clusters." *Proceedings of the National Academy of Sciences of the United States of America* 106 (9): 3282–3287.

Spellman, Paul T, and Gerald M Rubin. 2002. “Evidence for Large Domains of Similarly Expressed Genes in the Drosophila Genome.” *Journal of Biology* 1 (1): 5.

Splinter, Erik, and Wouter de Laat. 2011. “The Complex Transcription Regulatory Landscape of Our Genome: Control in Three Dimensions.” *The EMBO Journal* 30 (21): 4345–4355.

Sturgill, David, Yu Zhang, Michael Parisi, and Brian Oliver. 2007. “Demasculinization of X Chromosomes in the Drosophila Genus.” *Nature* 450 (7167): 238–242.

Sémon, Marie, and Laurent Duret. 2006. “Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals.” *Molecular Biology and Evolution* 23 (9): 1715–1723.

Tamames, Javier. 2001. “Evolution of Gene Order Conservation in Prokaryotes.” *Genome Biology* 2 (6): research0020.1–0020.11.

Vibranovski, Maria D, Hedibert F Lopes, Timothy L Karr, and Manyuan Long. 2009a. “Stage-specific Expression Profiling of Drosophila Spermatogenesis Suggests That Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-expressed Genes.” *PLoS Genetics* 5 (11): e1000731.

Vibranovski, Maria D, Yong Zhang, and Manyuan Long. 2009b. “General Gene Movement Off the X Chromosome in the Drosophila Genus.” *Genome Research* 19 (5): 897–903.

Wasbrough, Elizabeth R, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L Karr. 2010. “The *Drosophila melanogaster* Sperm proteome-II (DmSP-II).” *Journal of Proteomics* 73 (11): 2171–2185.

Weber, Claudia C, and Laurence D Hurst. 2011. “Support for Multiple Classes of Local Expression Clusters in *Drosophila melanogaster*, but No Evidence for Gene Order Conservation.” *Genome Biology* 12 (3): R23.

Woo, Yong H, Michael Walker, and Gary A Churchill. 2010. “Coordinated Expression Domains in Mammalian Genomes.” *PloS One* 5 (8): e12158.

Zhang, Yong E, Maria D Vibranovski, Benjamin H Krinsky, and Manyuan Long. 2010. “Age-dependent Chromosomal Distribution of Male-biased Genes in Drosophila.” *Genome Research* 20 (11): 1526–1533.

Chapter 5

Retrogene expression is associated with residence within testis gene neighbourhoods in *Drosophila melanogaster*

5.1 Introduction

Gene creation is important in the generation of biological complexity (Lynch and Conery 2003; Lynch 2002). New genes can be created *de novo*, by the fusion of existing genes or by gene duplication. Gene duplication can be DNA or RNA based, and both have significantly contributed to the evolution of genomic novelty. For example, of 59 young, essential genes identified in Drosophila, approximately 30% could be determined to be the result of RNA based duplication events and an equal amount (~30%) the result of DNA based duplication (Chen, Zhang, and Long 2010). Furthermore, many new gene duplicates have male-biased expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004; Marques et al. 2005; Kaessmann 2010). In *Drosophila melanogaster* many have acquired functions sperm (Dorus et al. 2008). Including the retrogene *k81*, which is believed to be involved in the formation of the telomere capping complex in sperm, as it specifically associates with other known capping proteins (HOAP and HIP) on the telomeres throughout spermatogenesis and until zygote mitosis (Dubruille et al. 2010); and a further 20 retrogenes that encode *D. melanogaster* sperm proteome components (see Chapter 3). As such we propose that retrotransposition is an important process in the creation of new genes associated with male reproduction.

In retrotransposition a processed mRNA copy of a transcribed gene undergoes reverse transcription and is inserted into the genome, often at a distance from the original (parental) gene. Due to the mechanism of retrotransposition, the resulting gene copy (retrogene) shares few genetic features of the original gene. The processing of the mRNA molecule results in the copy being intronless, and therefore it does not share the intron-exon structure of the original, rendering it unable to produce splice variants. In addition, the regulatory sequences of the parental gene are not copied unless they were located downstream of the transcriptional start site, thus new retrogenes must somehow acquire new promoter sequences if they are to be expressed (Kaessmann, Vinckenbosch, and Long 2009). Furthermore, as retrogenes are generally inserted at great distances from the parental gene, often inter-chromosomally, they do not share any higher-order regulation, such as chromatin domain structure, with their parental genes and as such are exposed to different selective regimes based upon the genomic architecture of their region of insertion. All of this means that the resulting retrogene is not

equivalent to the original gene and may not be immediately functional (Jun et al. 2009; Kaessmann, Vinckenbosch, and Long 2009).

Many studies have demonstrated a general tendency of retrogenes to have testis-specific, or testis biased, expression while their parental genes have wider tissue expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004; Bai et al. 2007; Langille and Clark 2007), or have demonstrated that many retrogenes have evolved novel functions in spermatogenesis (Bradley et al. 2004; Ding et al. 2010) and encoding proteins incorporated in mature spermatozoa (Dorus et al. 2008; Vemuganti, de Villena, and O'Brien 2010). These studies demonstrate the important role of retrotransposition in the evolution of male genes. However, due to the mechanism of retrotransposition the evolutionary process by which retrogenes initially acquire new functions remains unclear and highly debated. As retrogenes in general do not inherit *cis*-regulatory sequences from their parental genes, the process by which retrogenes acquire expression remains unclear. There are several, non-mutually exclusive, methods by which this may occur. It is possible that mutations in the nucleotide sequence close to the newly inserted retrogene could result in *de novo* creation of new promoter sequences (Bai, Casola, and Betrán 2009; Kaessmann, Vinckenbosch, and Long 2009). Consistent with this mechanism an enrichment of a novel motif was identified in upstream regions of testis expressed retrogenes (Bai, Casola, and Betrán 2009). An alternative mechanism is that retrogenes are able to recruit regulatory sequences through the sharing, or co-opting, of regulatory elements from neighbouring genes (Vinckenbosch, Dupanloup, and Kaessmann 2006; Bai, Casola, and Betrán 2008).

It is highly likely that only genes expressed in the germline should be able to produce heritable retrogenes, and has been proposed that retrogenes should be enriched in regions easily accessible in the genome, such as open chromatin domains (Kaessmann, Vinckenbosch, and Long 2009). This is supported by the observation that p-elements (a type of transposon specific to *D. melanogaster*) tend to insert into open chromatin regions in the germline of the sex they are in (Bownes 1990). In males this suggests that testis neighbourhoods may be likely places for retrogenes to insert. Several authors have commented on the potential importance of the site of retrogene insertion in determining a retrogenes fate (Betrán, Thornton, and Long 2002; Bai, Casola, and Betrán 2008; Dorus et al. 2008). Previous studies have identified an excess of retrogenes in testis gene neighbourhoods but not confirmed a correlation between expression of the retrogenes and their closest neighbours (Bai, Casola, and Betrán 2008). However, the study of nearest neighbours may not be appropriate, as testis gene neighbourhoods in *D. melanogaster* often span multiple genes (Boutanaev et al. 2002; see also Chapter 4) and there is evidence supporting their regulation by chromatin condensation and position in the nucleus (Shevelyov et

al. 2009). We propose that retrogene residence in these testis neighbourhoods may be higher than expected due to (1) insertional bias of retrogenes into these regions due to increased accessibility resulting from open chromatin conformation in testis, and/or (2) a bias in retention of retrogenes in these regions. Further we propose that there will be an association between retrogene testis expression and residence within testis gene neighbourhoods. This may be due to (1) the increased availability of regulatory regions to share or co-opt within co-expressed gene neighbourhoods (2) the permissive expression environment of open chromatin regions in the testis (Babushok, Ostertag, and Kazazian 2007) and/or (3) selection to maintain the integrity of the neighbourhood.

Here we investigated the extent to which the genomic landscape has affected the evolution of retrogene expression by determining whether there is an association between retrogene expression in the testis and residence within a testis neighbourhood. We observed a significant over-representation of retrogenes within testis neighbourhoods than expected by either random genomic distribution or an insertion bias. Further we observe a significant association between retrogene residence in a testis neighbourhood and testis-biased expression of the retrogene, with all retrogenes within neighbourhoods having testis-expression similar to that of the remaining genes in their neighbourhoods.

5.2 Materials and methods

5.2.1 Identification of retrogenes

Analyses were based on previously characterised *Drosophila melanogaster* retrogenes (Bai et al. 2007; Zhou et al. 2008). However, several retrogene-parent pairs were removed from our analyses, including three pairs due to the retrogenes currently being annotated as "pseudogene_attribute" in FlyBase (www.flybase.org). The original datasets also provided three pairs of genes where the original retrogene had subsequently undergone tandem duplication; we retained one of the retrogenes from each pair. This resulted in 96 retrogene-parent gene pairs in our analyses.

5.2.2 Retrogene, age and location

We obtained the chromosomal location of each retrogene and parental gene from FlyBase. In addition, for each retrogene we obtained its annotated orthologs throughout the Drosophila genus from FlyBase (gene_orthologs_fb_2012_04.tsv.gz; 24/07/2012). The presence/absence of annotated orthologs, for each retrogene, was used to parsimoniously determine the ancestral node in which it was first observed, and this was used as a proxy for retrogene age. We divided retrogenes into two classes: "young" and "old", based upon the absence or presence (respectively) of orthologs. Old retrogenes had an annotated ortholog in *D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis* and/or *D. grimshawi*, while young retrogenes did not have an annotated ortholog in any of these species. We used Fisher's two-tailed tests to determine whether the proportions of "old" retrogenes were significantly different between retrogenes in neighbourhoods and retrogenes outside of neighbourhoods.

5.2.3 Identification of testis gene neighbourhoods

We used previously curated expression data and documented testis neighbourhoods (see Chapter 4). Briefly, probe annotation data (Drosophila_2.na32.annot.csv) was downloaded from Affymetrix via FlyAtlas (www.flyatlas.org; 19/3/2012) (Chintapalli, Wang, and Dow 2007). Probes were removed if no alignment location was provided, probes were aligned to multiple genome locations, probes had no FlyBase symbol, FlyBase identifier or RefSeq identifier, probes were assigned to heterochromatic regions, chromosome 4 or chromosome U, or the designated gene had been withdrawn from FlyBase. Microarray expression data was obtained from FlyAtlas (3/4/2012) (Chintapalli, Wang, and Dow 2007). Locations of all annotated genes within the *D. melanogaster* genome were obtained from FlyBase (19/3/2012). Genes with matching start sites had their expression data averaged. Genes were ordered along their respective chromosome by start sites. A testis gene was defined as one with 2-fold higher expression in the testis compared to the remainder of the fly. This relative expression was determined by calculating the ratio of testis to whole fly expression. We defined a testis

neighbourhood as a genomic region that contained a minimum of 3 genes, a minimum of 66% genes within the region were defined as testis genes, and the neighbourhood was identified in both strand orientations (defined as a conserved loose neighbourhood in Chapter 4).

5.2.4 Assessment of the association between retrogene expression and location

We determined if an association between retrogene location and expression existed using Fisher's two-tailed exact test. To assess whether any association could be due to stochastic insertion or retention of retrogenes into testis neighbourhoods, we calculated the number of retrogenes expected to reside in testis neighbourhoods based on (1) the proportion of genes residing within testis neighbourhoods and (2) the proportion of the genome comprised of testis neighbourhoods. In order to determine if there was an insertional bias of retrogenes into testis neighbourhoods we obtained a list of all natural transposable elements (TE) sites in *D. melanogaster* from FlyMine (www.flymine.org; Template: organism/Natural transposable elements and their insertion sites; 1/8/2012) and determined the expected number of TE to reside in testis neighbourhoods based on the proportion of the genome comprised of testis neighbourhoods by size. We used Fishers' two-tailed exact tests to assess whether there were statistically significant differences between the observed and expected number of retrogenes.

5.2.5 Comparison of expression

In addition to expression information from FlyAtlas (Chintapalli, Wang, and Dow 2007), we also obtained expression data from samples enriched with cells in mitotic, meiotic and post-meiotic stages of spermatogenesis (Vibranovski et al. 2009). We used this information to determine the expressional range (defined as the range between the highest and lowest recorded expression of all genes within a specified genomic region) for each neighbourhood that contained a retrogene. We were then able to determine whether the retrogenes' expression was within the range of expression of its neighbourhood, or whether retrogenes were generally the highest or lowest expressed genes within a neighbourhood. Further, we compared the expression of (1) testis-expressed retrogenes within neighbourhoods to testis-expressed retrogenes outside of neighbourhoods and (2) compared the expression of genes not located within neighbourhoods, genes located within testis neighbourhoods that did not contain a retrogene and genes located within testis neighbourhoods with a resident retrogene. All comparisons we statistically assessed using Kolmogorov- Smirnov tests.

5.3 Results

5.3.1 Testis expressed retrogenes are associated with testis gene neighbourhoods

Our dataset contained 96 retrogene-parent gene pairs, including 3 instances in which a single parental gene (*CkIalpha, Cnx99A, Vha16-1*) had independently produced two retrogenes (*CG7094* and *CG2577*; *CG1924* and *CG9905*; *vha16-2* and *CG9013*, respectively). Two retrogenes (*RpS15Ab* and *CG13402*) did not have expression data and were conservatively classified as not testis expressed. Similarly their parental genes (*RpS15Aa* and *CG32601*, respectively) did not have expression data, and were therefore conservatively classified as not testis expressed. Consistent with previous studies the majority of retrogenes (63 out of 96) were considered to have testis-biased expression (hereafter termed testis retrogenes) while very few (9 out of 93) parental genes were considered to have testis-biased expression. We identified 30 retrogenes within 25 testis neighbourhoods, this did not include *CG7514*, which was in a conservative loose neighbourhood of 2 testis genes and was subsequently classified as "not clustered". We determined that there was a significant association between retrogene expression in the testis and residence within a testis neighbourhood ($p < 0.0001$). All retrogenes located within testis neighbourhoods were expressed at least 2-fold higher in the testis than the remainder of the fly. Of the remaining retrogenes, outside of neighbourhoods, 33 were testis expressed and 33 were not testis expressed. It is also noteworthy that only 2 parental genes (*CG32063* and *Prosβ5R1*) are located within testis neighbourhoods.

5.3.2 Testis neighbourhoods persist after the removal of associated retrogenes

To determine whether the identified testis neighbourhoods were an artefact of the retrogenes presence we removed all retrogenes within a neighbourhood and re-analysed that specific genome region to assess whether it, or a portion of it, remained a testis neighbourhood based on the parameters: a minimum of 3 genes and 66% testis genes. Of the 25 testis neighbourhoods, 19 persisted after the removal of the retrogenes, including four testis neighbourhoods that contained more than 1 retrogene. To confirm that this did not affect the association between retrogene expression in the testis and residence within a testis neighbourhood, we recalculated Fisher's exact p value after re-classifying the 6 retrogenes (those whose neighbourhoods did not exist without the retrogene) as not clustered. The association between retrogene testis expression and location remained significant ($p < 0.0001$).

5.3.3 There is an excess of retrogenes within testis neighbourhoods

We assessed whether the observed number of retrogenes within testis neighbourhoods could be explained by stochastic distribution of retrogenes. We therefore calculated the expected number of retrogenes to reside in testis neighbourhoods based on the proportion of the genome comprised of testis neighbourhoods. We observed significantly higher number of retrogenes within testis neighbourhoods (n = 30) than expected (n = 11; p = 0.0013) based upon the proportion of the genome comprising testis neighbourhoods by size (11.9%). We repeated this calculation using the proportion of genes within testis neighbourhoods (10.3%) to determine the expected random distribution of retrogenes. Again we found that a significantly higher number of retrogenes are located within testis neighbourhoods than expected (n = 10; p = 0.0006).

5.3.4 The excess of retrogenes residing in testis neighbourhoods is not due to an insertional bias

In order to assess whether the observed excess of retrogenes within testis neighbourhoods is due to an insertional bias, we determined whether a similar excess exists for TE. There are 4,851 natural TE insertion sites recorded in FlyMine (www.flymine.org) in the *D. melanogaster* genome, of which 514 are located within testis neighbourhoods. This is significantly lower than the number of TEs expected in testis neighbourhoods (n = 578; p = 0.0430) based upon the proportion of the genome constituting testis neighbourhoods by size (11.9%).

Table 5.1 Majority of retrogenes conform to the expressional range of their neighbourhood

Tissue	# Retrogenes within neighbourhood expressional range	Retrogenes with highest expression in neighbourhood	# Retrogenes with lowest expression in neighbourhood
Testis	22	*Cdlc2, CG9582, CG9254, CG10749. S-LAP7, tomboy40*	*robl22E, qjt*
Remainder of Fly	22	*Cdlc2, CG9582, CG9254, CG10749, S-LAP7*	*robl22E, qjt, vha16-2*
Mitosis	23	*Cdlc2, CG9254, CG7804, Rcd-1r, RpL37b, S-LAP7*	*robl22E*
Meiosis	24	*Cdlc2, CG9254, CG7804, tomboy40, S-LAP7*	*robl22E*
Post-meiosis	19	*Cdlc2, CG9582, CG9254, CG10749, ran-like, S-LAP7, tomboy40, Trxr-2*	*robl22E, CG10104, qjt*

5.3.5 Retrogene expression is consistent with the expression of the remaining neighbourhood genes

Retrogenes located within testis neighbourhoods have significantly higher expression in the testis (D = 0.2808, p = 0.027), but not in the remainder of the fly (D = 0.1483, p = 0.583), compared to the average expression of the remaining genes in these neighbourhoods (Figure 5.1a). Similarly, retrogenes residing in testis neighbourhoods have significantly higher average expression during mitosis (D = 0.3041, p = 0.012), meiosis (D = 0.3196, p = 0.007) and post-meiotic (D = 0.2646, p = 0.043) stages of spermatogenesis compared to the average expression of the remaining genes in the neighbourhoods (Figure 5.1b). However, this may be due to (1) a small number of very highly expressed retrogenes or (2) non-testis expressed retrogenes within the testis neighbourhoods. Therefore we determined whether the retrogenes within testis neighbourhoods had expression between the maximum and minimum levels of expression in their specific neighbourhood (expressional range of the neighbourhood). We observed, that the majority of retrogenes in testis neighbourhoods had expression values within the expressional

range of their neighbourhood for testis, whole fly and throughout spermatogenesis (Table 5.1). However in each tissue/stage a small number of retrogenes had the highest expression of all genes within their resident neighbourhood.

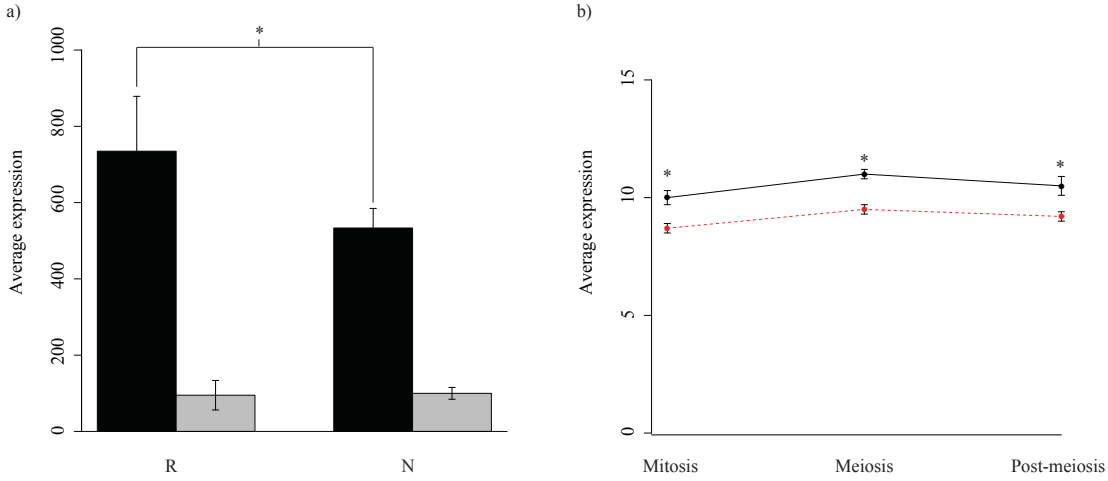

Figure 5.1 Comparison of the average expression of retrogenes within testis neighbourhoods and the remaining neighbourhood genes.

(a) Average expression in the testis (black) and remainder of the fly (grey) of retrogenes residing within neighbourhoods (R) and the remaining neighbourhood genes (N). (b) Average expression throughout spermatogenesis of retrogenes within neighbourhoods (solid, black line) and the remaining genes within those neighbourhoods (dashed, red line). Asterisks (*) indicate significant differenced between retrogene expression and the expression of other genes within those neighbourhoods, as determined by non-parametric Kolmogorov-Smirnov tests. Standard errors are provided.

5.3.6 Neighbourhoods have higher expression in male tissues than other genomic regions

We sought to determine if there were differences in expression within genomic regions that may explain why there is an excess of retrogenes within certain neighbourhoods. We removed retrogenes from this analysis and divided the remaining genes into three categories: (1) genes that are not found in testis neighbourhoods, (2) genes that are found in testis neighbourhoods that do not contain a retrogene and (3) genes that are found in testis neighbourhoods that have a resident retrogene. Predictably we found that genes outside of testis neighbourhoods had significantly lower average expression in the testis compared to genes found in testis neighbourhoods, regardless of whether these neighbourhoods contained retrogenes (Figure 5.2a). Similarly we found that genes outside of testis neighbourhoods has significantly lower average expression in all stages of spermatogenesis compared to genes in testis neighbourhoods, regardless of whether these neighbourhoods contained retrogenes (Figure 5.2b)

When we compared the expression of genes within neighbourhoods that contain a retrogene and those within neighbourhoods that do not contain a retrogene, we observed no significant difference in the average testis expression (D = 0.0897; p = 0.1360) (Figure 5.2a). Similarly we observed no significant difference in the average expression between genes that reside in neighbourhoods containing retrogenes and those that do not in the mitotic (D = 0.0977; p = 0.0962) and post-meiotic (D = 0.0540; p = 0.7430) stages of spermatogenesis; although we did observe a significant difference during meiosis (D = 0.1394; p = 0.0042) (Figure 5.2b)

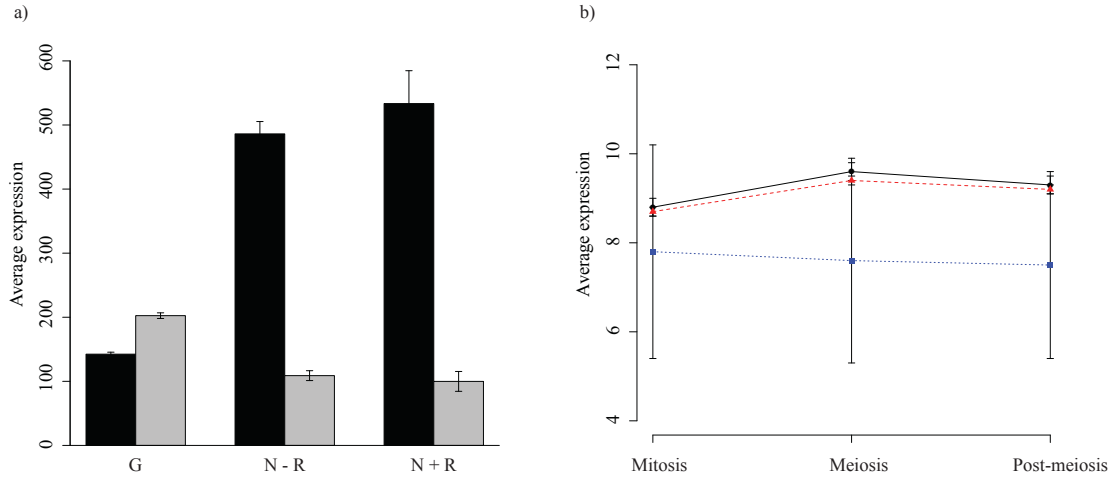

Figure 5.2 Comparison of expression between genes within neighbourhoods and those elsewhere.
(a) Average expression in the testis (black) and remainder of the fly (grey) for genes that are not found in testis neighbourhoods (G), genes found within neighbourhoods that do not contain retrogenes (N - R) and genes found within neighbourhoods that do contain retrogenes (N + R). (b) Average expression throughout spermatogenesis for genes not found in testis neighbourhoods (blue, squares), genes found within neighbourhoods that do not contain retrogenes (red, triangles) and genes found within neighbourhoods that do contain retrogenes (black, circles). Standard errors are provided for all average values.

5.3.7 Comparison of testis retrogenes residing in and out of neighbourhoods

We sought to determine whether there was any difference, beyond location, between testis-expressed retrogenes residing within testis neighbourhoods and those residing elsewhere in the genome. We compared the proportion of "old" testis retrogenes residing within neighbourhoods (17 out of 30; 56.7%) to the proportion of "old" testis retrogenes residing elsewhere (25 out of 33; 75.8%). These proportions were not significantly different (p = 0.1200). Similarly, we found that there was no significant difference in the proportion of "old" retrogenes that were

testis-expressed (42 out of 63; 66.7%) compared to the proportion of "old" retrogenes that were not testis-expressed (27 out of 33; 81.8%) (p = 0.1533). However, we did observe that the proportion of "old" retrogenes within testis neighbourhoods was significantly lower than the proportion of "old" retrogenes not residing within testis neighbourhoods (52 out of 66; 78.8%) (p = 0.0305). Further, we compared the expression of testis-expressed retrogenes found within testis neighbourhoods and testis-expressed retrogenes not located within a neighbourhood. We found no significant differences between the two sets of testis retrogenes in the testis (D = 0.1364, p = 0.912), the remainder of the fly (D = 0.1273, p = 0.947), or in the mitotic (D = 0.1667, p = 0.760), meiotic (D = 0.1333, p = 0.936) or post-meiotic (D = 0.1333, p = 0.936) stages of spermatogenesis (Figure 5. 3).

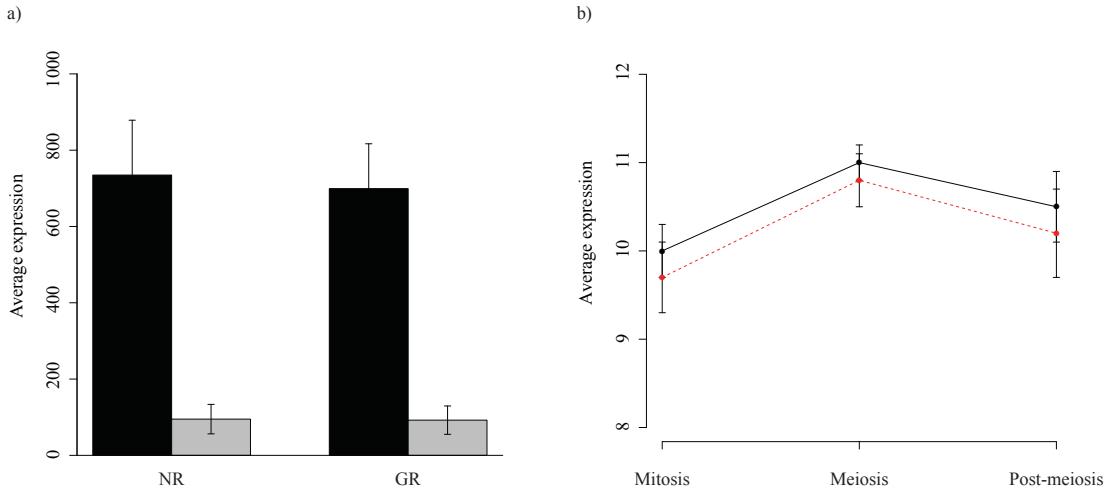

Figure 5.3 Comparison of expression of testis retrogenes residing in and out of testis neighbourhoods.

(a) Average expression in the testis (black) and the remainder of the fly (grey) of testis-expressed retrogenes residing within testis neighbourhoods (NR) and testis-expressed retrogenes located outside of testis neighbourhoods (GR). (b) Average expression throughout spermatogenesis of testis-expressed retrogenes residing within testis neighbourhoods (black, solid line) and testis-expressed retrogenes not located within testis neighbourhoods (red, dotted line). Standard errors are provided for all averages.

5.4 Discussion

It has been observed that retrotransposition is a "shotgun" approach, which scatters the gene copies throughout the genome, some of which will insert into "fertile" genomic regions and give rise to new genes, while others will not (Brosius 1991). As such it has been proposed that the location in which a retrogene resides is important in its evolution (Betrán, Thornton, and Long 2002; Bai, Casola, and Betrán 2008; Dorus et al. 2008). As many retrogenes acquire testis-biased expression (Betrán, Thornton, and Long 2002; Emerson et al. 2004) coupled with the hypothesis that genes expressed in the germline are more likely to produce heritable retrocopies and these copies are more likely to insert into accessible genomic regions, such as open chromatin regions containing genes expressed in the germline (Kaessmann, Vinckenbosch, and Long 2009), we proposed that genomic regions enriched in testis genes (i.e. testis gene neighbourhoods) may represent "fertile genomic regions" for retrogenes to insert. Furthermore, due to the potential for retrogenes to co-opt or share regulatory sequences (Vinckenbosch, Dupanloup, and Kaessmann 2006) and the selection for genes to conform to the expression pattern of these regions that this would influence retrogene expression. Consistent with these proposals we observed a significant excess of retrogenes within testis neighbourhoods and that there was a significant association between retrogene residence in a neighbourhood and the testis expression of the retrogene.

5.4.1 There is an enrichment of retrogenes in testis neighbourhoods

Previous studies addressing the question of how retrogene expression evolution relates to a retrogenes genomic location have either looked at individual, or small numbers, of retrogenes (Dorus et al. 2008) or have investigated the retrogene and a small number of its neighbouring genes (Bai, Casola, and Betrán 2008). In *D. melanogaster* Bai *et al*. (2008) investigated whether one or more of the four neighbouring genes surrounding a testis-expressed retrogene was likewise testis-expressed. They observed no correlation between retrogene testis expression and testis expression of a neighbouring gene (Bai, Casola, and Betrán 2008). However, we observed a significant excess of retrogenes residing within testis neighbourhoods. In addition we observed a significant association between retrogene testis expression and retrogene residence in a testis neighbourhood, which may explain the evolution of at least some retrogenes expression. These differences in results are likely due to differences in how we determined neighbouring genes (i.e. four neighbours vs. region enriched for testis expressed genes). We propose that our use of genomic regions enriched for genes expressed in the testis rather that an inflexible specified number of neighbours is more biologically realistic; as studies in both *Mus musculus* and *D. melanogaster* regarding the spatial organisation of testis genes have demonstrated that (1) there is significant co-localisation of these genes into gene neighbourhoods and (2) that these neighbourhoods are often larger than four genes, although the

range of neighbourhood sizes can be large (Boutanaev et al. 2002; Divina et al. 2005; Li, Lee, and Zhang 2005). Further support for the use of testis gene neighbourhoods rather than small numbers of neighbouring genes is the indirect evidence that these regions are in open chromatin conformation throughout the male germline, and therefore likely places for retrogenes to initially insert, as they have average higher expression in the testis and throughout spermatogenesis than the remainder of the genome (Figure 5.2).

The over-representation of retrogenes within testis neighbourhoods may be explained by either an insertional bias of retrogenes into these areas or a bias in the retention of retrogenes that insert into these genomic regions. Regarding an insertional bias, it is possible that testis neighbourhoods represent regions of the genome that are more amenable to retrogene insertion than other genomic regions, as in order to insert a new genomic element the DNA sequence must be accessible. In order for the DNA sequence to be accessible the genomic region must be in an open chromatin formation, which is indicative of transcriptionally active genes. As testis neighbourhoods are regions of the genome enriched for genes expressed in the testis and therefore likely to be required in the male germline, they will be in open chromatin regions at this time, and thus likely represent regions where insertion of new sequences is increased. An insertional bias may be assessed by the examination of the distribution of naturally occurring TE either due to the same mechanism of creation or as a by-product of the availability of sites amenable for insertion. Furthermore, the preference for insertion into open chromatin regions in the male germline has been experimentally investigated utilising p-elements. P-elements are TE that contain the bacterial gene that encodes *β*-galactosidase, as such the expression of *β*-galactosidase can indicate the location of insertion, and this indicated a preference for insertion into active germline locations (Bownes 1990). However, we observed significantly fewer TEs residing in testis neighbourhoods than expected by a random insertion model. Therefore we concluded that even if there was a preferential insertional bias for retrogenes into testis neighbourhoods, it could not explain the current disparities in the distributions of retrogenes and TEs. As such we propose that a retentional bias of retrogenes is a more likely explanation.

5.4.2 Retrogene presence in a testis neighbourhood effects expressional evolution

The proposal that there is a retention-based bias of retrogenes residing within testis neighbourhoods is further supported by the observation that all retrogenes within a testis neighbourhood are also testis-expressed, while only 50% of retrogenes elsewhere in the genome are similarly expressed. This suggests all retrogenes that persist in testis neighbourhoods have acquired testis-biased expression. Retrogenes that insert into testis neighbourhoods may be more likely to evolve testis expression for several reasons. Firstly, as spatially co-localised genes are often co-expressed and/or co-regulated (reviewed in Hurst, Pál, and Lercher 2004), it

is possible that the insertion of a new gene may disrupt the regulation of this neighbourhood. As such retrogenes that insert into gene neighbourhoods are likely to be under selection to avoid disruption of the neighbourhood, those that result in deleterious disruptions are likely to be purged from the genome. Second, and relatedly, retrogenes inserted into neighbourhoods are likely to be under selection to conform to the expression of the remainder of the neighbourhoods due to the likelihood that only during the periods in which the remaining genes are required (chromatin structure uncondensed) will the new gene be able to express itself and therefore become functional. Finally, retrogenes that insert into established gene neighbourhoods are likely to have similar expression to the remaining genes because it is only from these genes that retrogenes can co-opt or otherwise share regulatory sequences (Vinckenbosch, Dupanloup, and Kaessmann 2006; Dorus et al. 2008). The influence of the testis neighbourhood on retrogene expression is further highlighted by the observation that retrogene expression tends to be within the expressional range of the other genes within their neighbourhood. However, a small proportion of retrogenes are the most highly expressed gene within their neighbourhood, including two genes that encode proteins identified in mature *D. melanogaster* sperm (*Cdlc2* and *S-LAP7*) (Dorus et al. 2006; Dorus et al. 2008; Wasbrough et al. 2010; Dorus, Wilkin, and Karr 2011) and are continuously found to be the most highly expressed gene within their neighbourhoods (Table 5.1). We therefore suggest that these retrogenes may have important roles in spermatogenesis and related processes to have acquired such high testis-biased expression, and as such are candidates for further more targeted functional studies.

5.4.3 Genomic location cannot fully explain retrogene expression evolution

Despite our evidence that retrogene residence within testis neighbourhoods has substantially influenced the expression evolution of these retrogenes, genome location cannot completely explain how retrogenes acquire expression; as a large proportion of retrogenes outside of these neighbourhoods are also testis expressed. When we compared these testis retrogenes we found no differences in the age of retrogenes within these groups or in the patterns of their expression (Figure 5.3). The questions therefore remain as to how these retrogenes outside of testis neighbourhoods have gained expression, in particular testis expression, and why have half of these retrogenes become highly expressed in the testis but the remainder have not. However, when we considered all retrogenes, regardless of expression in the testis, we did observe a significantly higher proportion of retrogenes outside of testis neighbourhoods, compared to retrogenes within testis neighbourhoods, are old. It is possible that those testis retrogenes outside of testis neighbourhoods were originally resident in testis neighbourhoods, and have relocated away from their neighbourhood after acquiring testis expression. Alternatively these testis retrogenes outside of neighbourhoods may have been part of testis neighbourhoods that have not been conserved over evolutionary time. To determine whether testis retrogenes not in

neighbourhoods acquired testis expression while in ancestral testis neighbourhoods will require similar studies to be undertaken in sister Drosophila species to reconstruct the ancestral state.

5.4.4 Summary

Retrotransposition has produced many new genes that have testis expression and have evolved important functions in sperm and spermatogenesis. The mechanism of retrotransposition poses many questions regarding the evolution of retrogene function and expression. One of the greatest influences in the acquisition of retrogene expression is believed to be the genomic region into which it is inserted. Utilising identified genomic regions that are enriched for genes expressed in the testis (Chapter 4) we demonstrate that these regions contain an excess of retrogenes, which cannot be explained by random insertion of retrogenes or an insertional bias. Furthermore we found that all retrogenes within these testis neighbourhoods were expressed in the testis, and that there was a general tendency for the retrogenes to have similar expression to the remaining genes in their neighbourhood. As such we propose that the genomic region into which a retrogene is inserted is pivotal in the evolutionary retention and expression acquisition of that retrogene. However, as not all retrogenes that have acquired testis expression reside in indentified testis neighbourhoods, genome location is not the only mediator of retrogene evolution and its acquisition of functional expression.

5.5 References

Babushok, D V, E M Ostertag, and H H Kazazian. 2007. “Current Topics in Genome Evolution: Molecular Mechanisms of New Gene Formation.” *Cellular and Molecular Life Sciences* 64 (5): 542–554.

Bai, Yongsheng, Claudio Casola, and Esther Betrán. 2008. “Evolutionary Origin of Regulatory Regions of Retrogenes in Drosophila.” *BMC Genomics* 9: 241.

Bai, Yongsheng, Claudio Casola, and Esther Betrán. 2009. “Quality of Regulatory Elements in Drosophila Retrogenes.” *Genomics* 93 (1): 83–89.

Bai, Yongsheng, Claudio Casola, Cédric Feschotte, and Esther Betrán. 2007. “Comparative Genomics Reveals a Constant Rate of Origination and Convergent Acquisition of Functional Retrogenes in Drosophila.” *Genome Biology* 8 (1): R11.

Betrán, Esther, Kevin Thornton, and Manyuan Long. 2002. “Retroposed New Genes Out of the X in Drosophila.” *Genome Research* 12 (12): 1854–1859.

Boutanaev, Alexander M, Aila I Kalmykova, Yuri Y Shevelyov, and Nurminsky Dmitry I. 2002. “Large Clusters of Co-expressed Genes in the Drosophila Genome.” *Nature* 420 (6916): 666–669.

Bownes, M. 1990. “Preferential Insertion of P Elements into Genes Expressed in the Germ-line of Drosophila Melanogaster.” *Molecular & General Genetics* 222 (2-3): 457–460.

Bradley, Julie, Andrew Baltus, Helen Skaletsky, Morgan Royce-Tolland, Ken Dewar, and David C Page. 2004. “An X-to-autosome Retrogene Is Required for Spermatogenesis in Mice.” *Nature Genetics* 36 (8): 872–976.

Brosius, J. 1991. “Retroposons-Seeds of Evolution.” *Science* 251 (4995): 753.

Chen, Sidi, Yong E Zhang, and Manyuan Long. 2010. “New Genes in Drosophila Quickly Become Essential.” *Science* 330 (6011): 1682–1685.

Chintapalli, Venkateswara R, Jing Wang, and Julian A T Dow. 2007. “Using FlyAtlas to Identify Better Drosophila Melanogaster Models of Human Disease.” *Nature Genetics* 39 (6): 715–720.

Ding, Yun, Li Zhao, Shuang Yang, Yu Jiang, Yuan Chen, Ruoping Zhao, Yue Zhang, et al. 2010. “A Young Drosophila Duplicate Gene Plays Essential Roles in Spermatogenesis by Regulating Several Y-linked Male Fertility Genes.” *PLoS Genetics* 6 (12): e1001255.

Divina, Petr, Cestmír Vlcek, Petr Strnad, Václav Paces, and Jirí Forejt. 2005. “Global Transcriptome Analysis of the C57BL/6J Mouse Testis by SAGE: Evidence for Nonrandom Gene Order.” *BMC Genomics* 6: 29.

Dorus, Steve, Scott A Busby, Ursula Gerike, Jeffrey Shabanowitz, Donald F Hunt, and Timothy L Karr. 2006. "Genomic and Functional Evolution of the *Drosophila melanogaster* Sperm Proteome." *Nature Genetics* 38 (12): 1440–1445.

Dorus, Steve, Zoë N Freeman, Elizabeth R Parker, Benjamin D Heath, and Timothy L Karr. 2008. "Recent Origins of Sperm Genes in Drosophila." *Molecular Biology and Evolution* 25 (10): 2157–2166.

Dorus, Steve, Elaine C Wilkin, and Timothy L Karr. 2011. "Expansion and Functional Diversification of a Leucyl Aminopeptidase Family That Encodes the Major Protein Constituents of Drosophila Sperm." *BMC Genomics* 12: 177.

Dubruille, Raphaëlle, Guillermo A Orsi, Lætitia Delabaere, Elisabeth Cortier, Pierre Couble, Gabriel A B Marais, and Benjamin Loppin. 2010. "Specialization of a Drosophila Capping Protein Essential for the Protection of Sperm Telomeres." *Current Biology* 20 (23): 2090–2099.

Emerson, J J, Henrik Kaessmann, Esther Betrán, and Manyuan Long. 2004. "Extensive Gene Traffic on the Mammalian X Chromosome." *Science* 303 (5657): 537–540.

Hurst, Laurence D, Csaba Pál, and Martin J Lercher. 2004. "The Evolutionary Dynamics of Eukaryotic Gene Order." *Nature Reviews. Genetics* 5 (4): 299–310.

Jun, Jin, Paul Ryvkin, Edward Hemphill, Ion Mandoiu, and Craig Nelson. 2009. "The Birth of New Genes by RNA- and DNA-mediated Duplication During Mammalian Evolution." *Journal of Computational Biology* 16 (10): 1429–1444.

Kaessmann, Henrik. 2010. "Origins, Evolution, and Phenotypic Impact of New Genes." *Genome Research* 20 (10): 1313–1326.

Kaessmann, Henrik, Nicolas Vinckenbosch, and Manyuan Long. 2009. "RNA-based Gene Duplication: Mechanistic and Evolutionary Insights." *Nature Reviews. Genetics* 10 (1): 19–31.

Langille, Morgan G I, and Denise V Clark. 2007. "Parent Genes of Retrotransposition-generated Gene Duplicates in *Drosophila melanogaster* Have Distinct Expression Profiles." *Genomics* 90 (3): 334–343.

Li, Quan, Bernett T K Lee, and Louxin Zhang. 2005. "Genome-scale Analysis of Positional Clustering of Mouse Testis-specific Genes." *BMC Genomics* 6: 7.

Lynch, Michael. 2002. "Gene Duplication and Evolution." *Science* 297 (5583): 945–947.

Lynch, Michael, and John S Conery. 2003. "The Origins of Genome Complexity." *Science* 302 (5649): 1401–1404.

Marques, Ana Claudia, Isabelle Dupanloup, Nicolas Vinckenbosch, Alexandre Reymond, and Henrik Kaessmann. 2005. “Emergence of Young Human Genes After a Burst of Retroposition in Primates.” *PLoS Biology* 3 (11): 1970–1979.

Shevelyov, Y Y, S A Lavrov, L M Mikhaylova, I D Nurminsky, R J Kulathinal, K S Egorova, Y M Rozovsky, and D I Nurminsky. 2009. “The B-type Lamin Is Required for Somatic Repression of Testis-specific Gene Clusters.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (9): 3282–3287.

Vemuganti, Soumya A, Fernando Pardo-Manuel de Villena, and Deborah A O’Brien. 2010. “Frequent and Recent Retrotransposition of Orthologous Genes Plays a Role in the Evolution of Sperm Glycolytic Enzymes.” *BMC Genomics* 11: 285.

Vibranovski, Maria D, Hedibert F Lopes, Timothy L Karr, and Manyuan Long. 2009. “Stage-specific Expression Profiling of Drosophila Spermatogenesis Suggests That Meiotic Sex Chromosome Inactivation Drives Genomic Relocation of Testis-expressed Genes.” *PLoS Genetics* 5 (11): e1000731.

Vinckenbosch, Nicolas, Isabelle Dupanloup, and Henrik Kaessmann. 2006. “Evolutionary Fate of Retroposed Gene Copies in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (9): 3220–3225.

Wasbrough, Elizabeth R, Steve Dorus, Svenja Hester, Julie Howard-Murkin, Kathryn Lilley, Elaine Wilkin, Ashoka Polpitiya, Konstantinos Petritis, and Timothy L Karr. 2010. “The *Drosophila melanogaster* Sperm proteome-II (DmSP-II).” *Journal of Proteomics* 73 (11): 2171–2185.

Zhou, Qi, Guojie Zhang, Yue Zhang, Shiyu Xu, Ruoping Zhao, Zubing Zhan, Xin Li, Yun Ding, Shuang Yang, and Wen Wang. 2008. “On the Origin of New Genes in Drosophila.” *Genome Research* 18 (9): 1446–1455.

Chapter 6

Conclusions and future work

The generation of large-scale data on proteomics, transcriptomics and genomics has provided us with extensive information with which to address questions surrounding the evolution of male reproductive genetics. In particular the technology to perform whole cell mass spectrometry on spermatozoa has provided previously inaccessible knowledge of sperm protein composition and therefore the evolution of sperm proteins. The goals of this dissertation were (1) to utilise a combined proteomic, genomic and transcriptomic approach to compare the sperm proteomes from two taxa with different sperm under different post-copulatory selective pressures, (2) explore the role of retrotransposition in the creation of novel sperm genes in these taxa and (3) investigate the role of genomic location on retrogene expression. In the remaining sections of this chapter we will highlight the major results from each previous chapter and briefly discuss how they have informed either the study of spermatozoa proteome evolution or the evolution of newly created male-biased genes, and suggest how these findings may influence the direction of future studies

In Chapter 2 we present the first detailed comparison of sperm proteomes from two different species: *Mus musculus* (mouse) and *Drosophila melanogaster*. In this analysis we discovered the extent to which the functional protein composition of the sperm proteome is conserved, highlighting the possibility that *D. melanogaster,* currently a model species for many mammalian processes, may be a useful tool in the study of genes involved in reproduction. In particular the use of Drosophila to study proteins involved in reproduction would be an improvement due to the plethora of available mutant strains and its faster generation time compared to mouse. Further, similarities in both protein domains and functions allowed us to identify a number of candidate genes that may have a role in insect fertilisation; a process in which few of the proteins involved are known. However, while the results of this study were informative, we hope that this study will encourage other similar studies among more closely related species, as this will provided nuanced information that is difficult to obtain when two species are so distantly related, such as the patterns of gene loss and gain.

The study of the sperm proteome composition highlighted a role for gene duplication in the creation of new sperm proteins. Many previous studies have highlighted the tendency for retrogenes to have male biased expression or to evolve novel functions that may be important to

spermatogenesis or in sperm fitness. However, no single study had combined a comparative approach with a detailed survey of how retrotransposition may have impacted sperm proteome evolution. In Chapter 3 we again employed a comparative approach to the study of the contribution retrotransposition has made to sperm evolution. We demonstrated the importance of retrotransposition to sperm proteome evolution, by the contribution of novel proteins to the sperm proteome of both Drosophila and mammals. Furthermore, we observed that retrogenes can act as markers of adaptive evolution. We determined that although retrotransposition had created sperm genes with metabolic functions in both mammals and Drosophila, the exact metabolic functions that were targeted were those underlying sperm fitness and therefore targets of post-copulatory selection. Further studies will be needed to determine whether other mechanisms of gene duplication, such as tandem gene duplicates, have a similar role in sperm proteome evolution. This study highlighted the potential that gene duplication patterns, both in terms of gene loss/gain and the functions they perform, can be used as markers of molecular evolution. However, the study of retrogenes raised the question: why do retrogenes acquire male biased functions? One of the prevalent hypotheses regarding retrogene evolution is the importance of the gene region into which it is inserted. Therefore in order to address whether genomic location is a principal factor in the expression evolution of retrogenes, we first needed to understand how the genome is organised.

Evidence suggests that the eukaryotic genome is organised based on the regulation of gene expression. Many studies have observed the spatial positioning of genes into regions where there is an enrichment of genes that are expressed at similar times and in similar tissues. Therefore in order to understand the linear organisation of genes within chromosomes we developed an effective new model for the identification of gene neighbourhoods (Chapter 4), which we applied to the identification of testis gene neighbourhoods. Consistent with previous methods we observed that there was significant co-localisation of testis genes. In addition we demonstrated that observations on the differences in gene content between the X chromosome and the autosomes can be dependent upon how researchers define a "testis gene", an important consideration that needs to be taken into account in future studies. Furthermore we observed that there was a large range in gene neighbourhood sizes. It is possible that this reflects the fact that chromatin domains are dynamic and do not regulate a set number of genes or genomic distance. However, although this model aids in the understanding of linear genome organisation, a complete understanding of gene position evolution and organisation is unlikely to occur using 2-dimensional models. We believe that future approaches need to combine the results of studies, such as those in Chapter 4, and reconcile them with 3-dimensional data including the position of chromosomes within the nucleolus, the resulting proximity of the identified gene neighbourhoods and the conformation of the chromatin structure in these

regions. Only by combining all three perspectives: gene neighbourhoods, chromatin conformation and chromosomal positioning, can we understand how gene expression is regulated and only then understand how genome organisation evolves.

Employing the results of the above model, in Chapter 5, we were able to demonstrate an excess of *D. melanogaster* retrogenes residing in testis gene neighbourhoods. Further we identified that all retrogenes within a testis neighbourhood were similarly testis expressed, providing compelling evidence for the important role of genome location on retrogene expressional evolution. In addition, we provided evidence that this enrichment is not due to an insertional bias, but is likely the result of selective retention of retrogenes. While it is unlikely that genome location is the only driver of retrogene evolution it appears to be a particularly influential factor. It will be interesting to see if this association between retrogene testis expression and residence in a testis gene neighbourhood is repeatable in other species. In particular it would be interesting to determine whether the co-localisation analysis (Chapter 4) and the retrogene association analysis (Chapter 5) produce comparable results when repeated in *M. musculus*.

This dissertation used a combination of genomic, proteomic and transcriptomic data in order to explore the evolution of male related genes. The primary focus of our research was on retrogenes, which have been intensely studied in mammals and insects and have been found to have a tendency to be male-biased in expression and to redistribute away from the X chromosome. In addition individual case studies have found evidence that retrogenes are involved in spermatogenesis and in producing transcripts incorporated into mature spermatozoa. We approached retrogene evolution from two directions (1) determining the pathways retrogenes function in, in spermatozoa, and (2) investigating whether the genomic location into which a retrogene inserts may effect their expression, and observed that selection acts at multiple levels: retrogenes tend to occur in metabolic processes associated with pathways under post-copulatory selection and that there is an association between retrogenes genomic location and testis expression. Finally, our study of retrogene evolution required us to develop a novel model for the identification of gene neighbourhoods. While this method allowed us to begin to answer questions about the association between retrogene genomic location and retrogene expression in the testis, it also addressed problems with previous identification methods and has resulted in a useful bioinformatic tool that can be applied to a wide variety of biological investigations, including changes in gene expression through different stages of cancer. We hope that future studies will also employ similar multi-targeted approaches as such studies provide a greater depth of information than individual isolated studies.

Appendix V

Table 1. Mammalian sperm retrogenes and their parental genes

Phylogenetic distribution	Sperm Retrogene*		X to autosome	Parental gene*	
	Gene symbol	Metabolic process (GO:0008152)		Gene symbol	Sperm proteome
Human	*PPP3R2*	Yes	No	*PPP3R1*	No
	PRKACG	Yes	No	*PRKACA*	Yes
Mouse	*Acot10*	Yes	Yes	*Acot9*	Yes
	Csl	Yes	No	*Cs*	Yes
	Aldoart1	Yes	No	*Aldoa*	Yes
	G6pd2	Yes	Yes	*G6pdx*	No
Rat	*RSA-14-44*	No	No	*Rhoa*	Yes
	RGD1560350	Yes	No	*Psma6*	Yes
	Aldoal1	Yes	No	*Aldoa*	Yes
	Prkar1a	Yes	No	*Prkar1b*	Yes
	LOC365778	No	No	*Fundc2*	No
Rodent	*1700071K01Rik*	No	No	*Phb*	Yes
	Gykl1	Yes	Yes	*Gyk*	Yes
	Mcts2	Yes	Yes	*Mcts1*	No
	1700029P11Rik	No	Yes	*Ndufb11*	Yes
	Pbp2	No	No	*Pebp1*	Yes
Euarchontoglire	*Gk2*	Yes	Yes	*Gyk*	Yes
	Pdha2	Yes	Yes	*Pdha1*	Yes
	Prps1l1	Yes	Yes	*Prps1*	Yes
	Rhob	No	No	*Rhoc*	Yes

Table 1 (continued)

Eutherian	*Actbl2*	No	No	*Actb*	Yes
	Dnajb3	Yes	No	*Dnajb6*	Yes
	Ftmt	Yes	No	*Fth1*	No
	Pgk2[†]	Yes	Yes	*Pgk1*	Yes
	Cetn1	Yes	Yes	*Cetn2*	No

* Identification in purified sperm by LC-MS/MS (Baker et al 2008; Baker et al. 2008; Baker et al. 2007; Cao, Gerton, and Moss 2006; Dorus et al. 2006; Dorus et al. 2010; Stein et al. 2006; Wasbrough et al. 2010)

[†] Mouse phenotypes include male infertility and asthenozoospermia (Danshina et al. 2010).

Appendix VI

Table 1. *D. melanogaster* sperm retrogenes and their parental genes

Phylogenetic distribution	Sperm retrogenes*		X to autosome	Parental genes*	
	Retrogene symbol	Metabolic process (GO:0008152)		Gene symbol	Sperm component
Sophophora subgenus	*CG17856*	Yes	Yes	*CG3560*	no
	CG5265	Yes	No	*CG1041*	no
	CG7514	No	No	*CG1907*	no
	Prosα6T	Yes	No	*Pros35*	no
Drosophilidae	*CG4706*	Yes	No	*Acon*	Yes
	S-LAP7	Yes	No	*S-LAP3*	Yes
	CG5718	Yes	No	*Scs-fp*	Yes
	Act87E	No	No	*Act57B*	No
	CG14508	Yes	No	*CG4769*	No
	Cdlc2	No	Yes	*Ctp*	No
	Ef1α48D	Yes	No	*Ef1α100E*	No
	Pglym87	Yes	No	*Pglym78*	No
	CG6255	Yes	No	*Scsα*	No
	Vha36	No	Yes	*CG8310*	No
	Hsp60B†	Yes	Yes	*Hsp60*	No
	Gskt (mojoless)†	Yes	Yes	*Sgg*	No
	CG10749	Yes	No	*CG7998*	No

Table 1 (continued)

Diptera	*Hsc70-4*	Yes	No	*Hsc70-1*	No
	CG11913	Yes	No	*CG1970*	No
	CG6180	No	No	*CG17919*	No

* Identification in purified sperm by LC-MS/MS (Wasbrough et al. 2010; Dorus et al. 2006)

† Male sterile alleles have been characterized for these genes (www.flybase.org).